61 (052) - 856

31 December 1965

Annual summary scientific report

SPEECH QUALITY MEASUREMENTS

Contractor: E. P. Hirschwehr

Principal Investigator: E. H. Rothauser
G. E. Urbanek

Cooperator: R. Eier
W. Pachl
F. Hezina

Distribution of this document is unlimited.

INSTITUT FÜR NIEDERFREQUENZTECHNIK
der Technischen Hochschule Wien
Vienna, Austria

The research reported in this document has been sponsored by the
UNITED STATES GOVERNMENT
under Contract 61 (052) - 856

AFCRL-66-170

61 (052) - 856

31 December 1965

Annual summary scientific report

SPEECH QUALITY MEASUREMENTS

Contractor : E. P. Hirschwehr

Principal Investigator : E.H. Rothauser
G.E. Urbanek

Cooperator : R. Eier
W. Pachl
F. Hezina

Distribution of this document is unlimited.

INSTITUT FÜR NIEDERFREQUENZTECHNIK
der Technischen Hochschule Wien
Vienna, Austria

The research reported in this document has been sponsored by the
UNITED STATES GOVERNMENT
under Contract 61 (052) - 856

TABLE OF CONTENTS

ABSTRACT

This paper is concerned with studies about the modified isopreference method for rating speech communication systems in view of speech quality. The concept of speech quality is studied by subjective measurements in terms of intelligibility and "preference". Listening experiments using the forced pair comparison technique have been performed with trained and untrained groups of listeners. Various kinds of speech signals from different systems have been compared with three idealized reference signals using noise in additive and multiplicative form as degradation signals. Different kinds of tests for preference, intelligibility, rank ordering and loudness are reported which were utilized to study several aspects of speech quality.

1. THE CONCEPT OF SPEECH QUALITY

1.1 Introduction

During the design, development and testing of systems for transmission, reproduction and artificial composition of speech signals there is a need for evaluation and for optimization criteria.

In the past intelligibility has been utilized as the main criterion for the evaluation of speech communication systems. During the last years modern speech processing techniques have reached a state of high perfection. Frequently, the intelligibility of the output speech signals of such systems is now so close to 100 % that intelligibility alone cannot suffice as a design criterion. In these cases one has to consider the full concept of "speech quality" rather than the aspect of intelligibility alone.

The measurement of the physical properties of a speech processing system and the combination of the results of such measurements in order to form a basis for comparing different such systems is at the moment only hopeful for systems with properties close to those of linear four-pole systems. Today for all more complex systems this objective approach is not feasible. At the present state of the art subjective measurements are necessary to find answers to the central questions : "How well does an average listener understand speech signals which are transmitted or created by the system under test" and "How does he like these speech signals, or the corresponding system, as a source of information?" The first question can be answered by intelligibility tests and the second by "preference" tests. While intelligibility tests are already a relatively well known tool, the additional evaluation of "preference" until now remained an only partially solved problem. "Preference" tests shall allow to express their aspect of speech quality in terms of a set of known standard reference signals or in terms of a continuously degradable reference signal. The reduction of the problem to only two key questions is an important constraint. Hopefully it allows to limit sufficiently the scope of the present work. It excludes the complex problems of speaker recognition and two-way communications.

The aim of the present study is to extend the knowledge around the concept of speech quality by the performance of subjective measurements. It concentrates on methods for preference testing and on their evaluation. A sufficient body of experimental data is being collected which will help to find a suitable method for preference testing and will show its possibilities and limitations. If possible a standard test procedure shall be proposed which allows to grade speech signals and permits meaningful comparisons between different types of systems and for comparisons between measurement results from different locations.

The scope of work described above is planned to be covered by end of 1966. The present interim report can therefore not provide answers to all of the problems in question. It describes the methods chosen for closer study and summarizes significant findings of the past year. In some cases the collected body of data was found to be still too small and did not allow for the conclusive determination of typical averages. The collection and evaluation of the additionally required data may necessitate some changes of statements in the present report. We hope that in the final report these changes will only be affirmations of our present views.

1.2 Speech Quality

Speech quality has many aspects. The degradation or loss of "quality" in transmitting speech over a telephone system may be seen completely different from the degradations in a vocoder system or some other speech processing device. In the first case it seems essential to evaluate the degree to which significant characteristics of the input signal are preserved by the communication channel. In the second case terms like "identification of the speaker" or the "emotional content of a message" may lose their importance and even their implication, e.g. one might device a transmission system with a "synthetic voice" which sounds natural compared with a typical human speaker but which surpresses all characteristics of the original speaker. Speech quality may be viewed to be a combination of the different attributes of speech signals which have to be preserved in order to give a listener the impression of a "high fidelity" system. It should describe the impression of an average listener when he compares a speech signal to speech patterns stored in his memory.

Speech quality includes various factors such as optimum loudness, timbre and rhythmic character, annoyance, a possible fatigue of the listener, speaker identifiability, naturalness, clarity, systematic amplitude or time distortions and many others. A quantitative definition of these factors is often not only difficult but sometimes next to impossible. This means that a detailed concept of speech quality to a certain extent depends upon interpretation, at least as long as there exists neither a comprehensive, accurate, and commonly recognized definition nor a standardized measurement procedure.

Our working-definition of "quality" contains besides intelligibility only a parameter called "preference". This term shall be an expression for the average attitude of a listener towards a test signal while comparing it consecutively with a reference speech signal with reproducible characteristics. Preference is thus a relative measure of quality.

1.3 Measurement Procedures

Direct measurements of the physical properties of a system for speech communication or speech processing may often be performed easily, but, as already mentioned, the results cannot always be related to the total subjective impression of an average listener. As quality is a psychological factor of speech communication, it requires, at least today, also psychological measurement techniques. The non-existence of a single listener with "average" properties and reactions to audio signals makes it necessary, in order to get statistical significance, to evaluate subjective judgements from a number of listeners. Unavoidably this implies a limitation of the expectable accuracy and reproducibility of the obtained results.

Speech signals with very different qualities may be rated by simple category tests. Here the listeners are classifying the test signal into a limited number of categories guided only by their personal memory and judgement. Higher reliability will result in another approach where the test signal is presented in pairs together with samples from a set of reference signals which represent the different categories. In such a procedure either the test signal or the reference signal, or both may be variable, and may exchange their relative position in the presented signal pairs. A summary of methods for assessing subjective factors of speech signals and a bibliography of work done before 1962 is contained in the paper of MUNSON and KARLIN /1/. The paper is concerned with a forced pair-comparison technique which is called isopreference method. Both, reference signal and test signal are varied. The reference signal was the voice of a real speaker or a hifi-tape recording of a speaker degraded by additive random noise. The results of this method are normally shown in the form of isopreference contours in a speech level versus noise level diagram (Fig. 1.1).

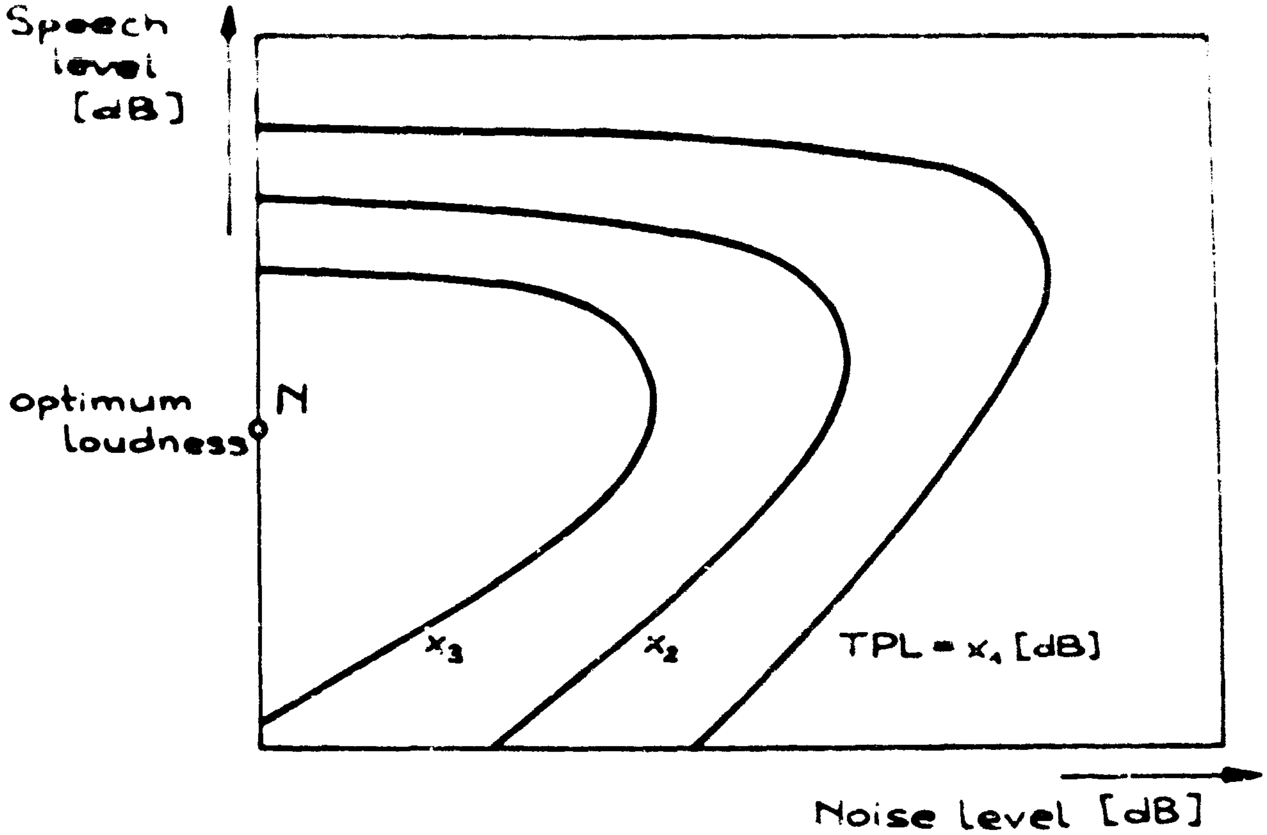


Fig. 1.1

The curves enclose the point or area N which represents the optimum setting of the test system with regards to the best adjustment of loudness and noice level. The method yields a quality rating in form of the "Transmission Preference Level" describing the isopreferent setting of the reference signal and additionally the optimum loudness level for the output of the test system.

1.4 The Proposed Method of Preference Testing

ROTHAUSER /2/ tried to duplicate some of the tests described by MUNSON and KARLIN. He had to find that the scattering of the test results was worse than anticipated. Presentation of successive test conditions along an isopreference contour showed suitable results because the test persons have only to cling to their specific criteria for preference judgements. But the deviations grow intolerably high when points on an isopreference contour with very low and very high levels of the test signal are compared. Here the judgements become very inconsistent because most of the listeners are annoyed by the unexpected and sometimes painfully high levels of the second signal. This means that the decisions of the listeners are influenced by the loudness levels of the previously presented signal pairs. The same accomodation effect can be observed for abrupt changes of the additive noise, which accompanies the test signal according to Fig. 1.1. In order to reduce the influence of these practically non-controlable conditioning effects the number of variables during a test run has been reduced as far as possible. Expressed in terms of Fig. 1.1 only the transmission preference level for the point N is determined. The loudness level of both test and reference signal are kept constant at a value equal to the optimum loudness of the special system. For a given test run only the S/N ratio of the reference signal is varied.

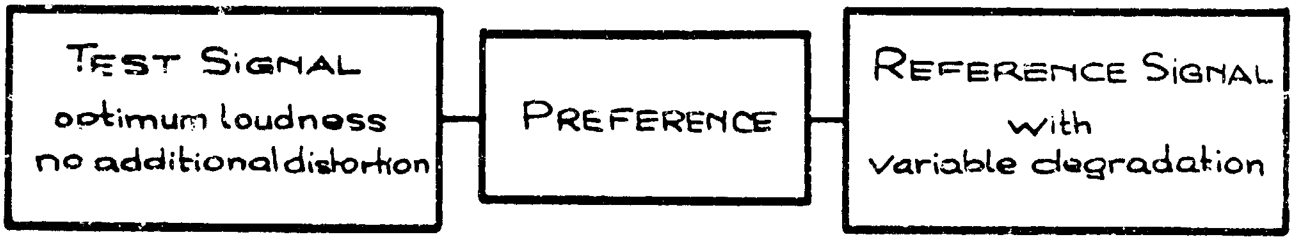


Fig. 1.2

Fig. 1.2 shows the variables in the modified preference test. The modified method yields not only a simplification of preference tests but also a substantial improvement with regards to accuracy and reproducibility of the test results.

2. PREFERENCE TESTING

2.1 Test Method

2.11 Description

The basic requirements for a measuring procedure are simplicity and reproducibility at different locations. Preference tests require only comparisons for a string of signal pairs, the test signal and the variable reference signal. In order to increase the accuracy the reference signal is presented to the listeners immediately before or after the test signal. The tests are not based on a particular aspect of quality but on overall preference with no requirement that the listeners have to categorize or to explain the reasons for their decisions. The listeners are not allowed to be indifferent in their decision between the two signals of any pair. They have to express their preference for one speech sample of each pair which they would prefer as a source of information. Preference is expressed in terms of a reference signal the quality of which is continously and reproducibly adjustable. The quality of the reference signal is degraded by adding a certain amount of a distortion signal to a hifi speech signal. Now the preference level of a test signal can be defined in terms of the S/N ratio of the reference signal where 50 % of all listeners favor the reference signal.

In order to avoid the difficulties MUNSON and KARLIN must have encountered by using a real speaker to produce the reference signal during the tests, only a hifi recording of such a speaker was used for the generation of the reference signal. Fig. 2.1 shows a simplified blockdiagram of the test set-up.

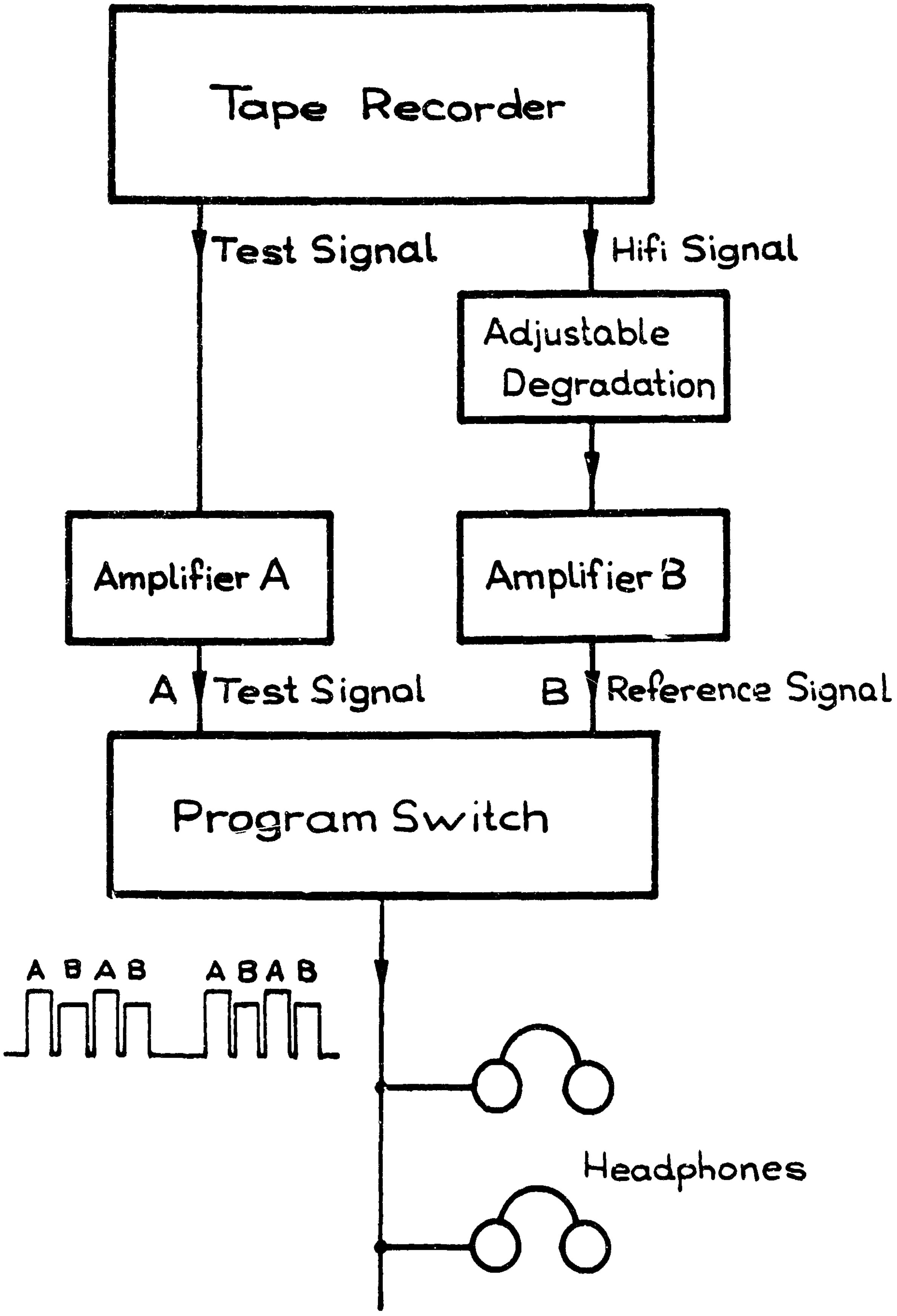

Tape Recorder
Test Signal
Hifi Signal
Adjustable Degradation
Amplifier A
Amplifier B
A
Test Signal
B
Reference Signal
Program Switch
A B A B A B A B
Headphones

Fig. 2.1

The test signal is recorded on one track of a stereo tape recorder, amplified and periodically fed to the receivers used by listeners. On the second track of the tape recorder a hifi speech signal is recorded. This signal can be reproducibly degraded for the generation of the reference signal. The test signal A and the reference signal B are presented in a successive and repetitive order to both ears of each listener via earphones. The speech level of both signals is adjusted to the respective optimum loudness for the particular signal which has to be determined also subjectively in a preparatory session.

A preference testing session may consist of about 5 test runs each consisting of approximately 15 pair comparisons. The test material is presented to the listeners in repeated signal pairs ordered as ABAB. Fig. 2.2 shows the mode of presentation by illustrating the time pattern and the variation of the S/N ratio of the reference signal B.

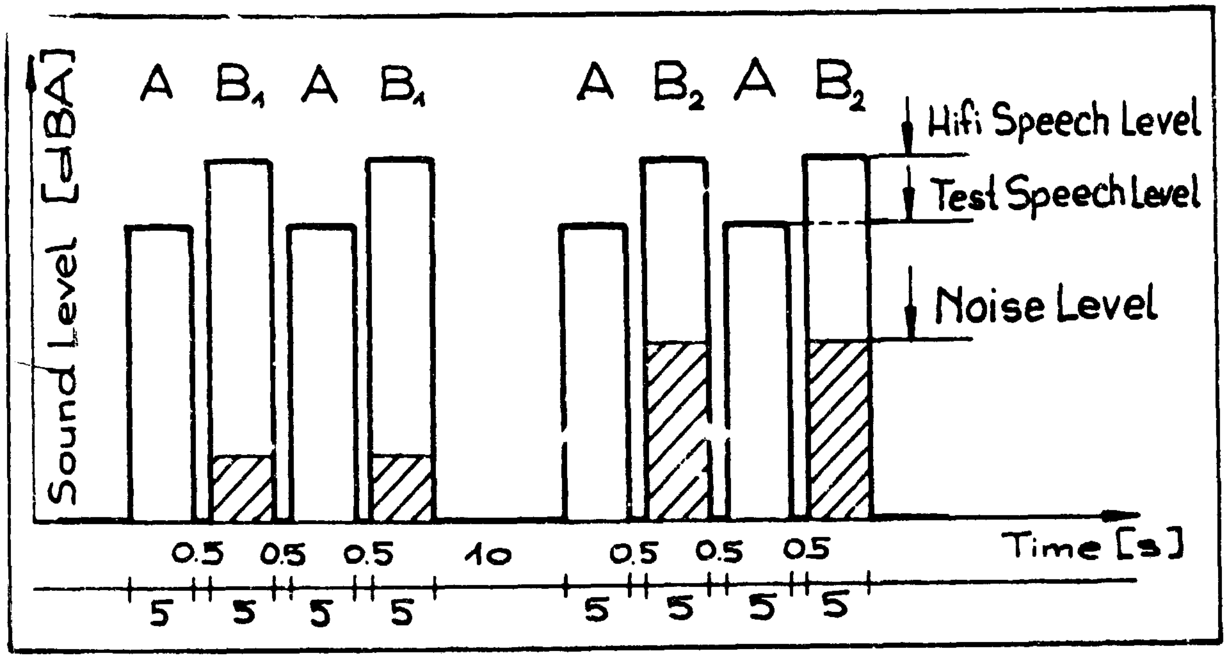


Fig. 2.2

The cross-hatching illustrates the constant amount of the S/N ratio of the reference signal for one repeated signal pair and the random variation for a consecutive pair. The duration of both speech samples A and B has been fixed 5 seconds and the interval between adjacent signal samples to 0.5 seconds. During a pause of 10 seconds between each of the repeated signal pairs, the listeners have to make and to indicate their decisions, and the operator is able to change the settings for the next pair comparison.

Normally for the evaluation of a test signal a preliminary test run is executed in order to establish the approximate value of the preference level and to determine the lower and upper limits of the S/N ratio for signal B, at which all listeners prefer either signal A or signal B. During the main test the incremental steps of the S/N ratio for signal B, i.e. the degradation of the reference signal are chosen much smaller. Then they should be small enough to cause inconsistent decisions by some listeners in the vicinity of their respective preference levels in order to get the highest possible accuracy in determining isopreference level.

2.2 Requirements

In spite of the drastic reduction of the variables during a single test run, as compared to the procedure followed by MUNSON and KARLIN, there remains a considerable number of parameters which may influence the results of a preference test. Table 2.1 lists the most important of these parameters. Even a rough estimate shows that there are more than thousand test conditions which differ in at least one of these parameters.

It has been one of the first tasks in the work reported here to select the most interesting and important test conditions. The specially marked parameters have been actually used in our tests. The present study is only concerned with continuous speech. It was decided to present both speech samples, i.e. the test signal and the hifi component of the reference signal, at optimum loudness levels which have been determined subjectively by the same listeners in previous sessions.

For the presentation of all signals headphones were chosen in order to avoid the difficulties which occur in conjunction with loudspeakers and acoustics. It may be possible that for later investigations also loudspeaker presentations will become more interesting in conjunction with the testing of speech signals under ambient noise conditions.

2.21 Reference Signals

The selection of the "best" reference signal for the purposes of preference testing is not easy and necessitates some compromises.

REFERENCE SIGNAL

TEXT :	continuous ○	words	syllables
LOUDNESS :	variable adjusted	fixed	optimum ○
PRODUCED BY :	live transmission system	idealized system	
		- hifi + noise ○	
		- hifi x (1 + k noise) ○	
ADDITIONAL		- real speaker + noise	
DISTORTIONS :	yes — no		
	telephone	vocoder	pulsedeltamod

TEST SIGNAL

TEXT :	continuous ○	words	syllables
LOUDNESS :	variable adjusted	fixed	optimum ○
PRODUCED BY :	live transmission system	idealized system	
		- hifi + noise ○	
		- hifi x (1 + k noise) ○	
		- real speaker + noise	

telephone ○ vocoder ○ pulsedeltamod ○ any new system ○

MODE OF PRESENTATION

TRANSDUCER :	headphones ○	handsets	loudspeakers
AMBIENT NOISE :	none ○	produced by loudspeakers	

- natural
  - office noise
  - typical noises
- artificial
  - wide band
  - band shaped

PRESENTATION FRAME

LISTENING GROUP

SIZE :	small (8 - 10) ○	large (> 50) ○
TRAINING :	trained	untrained
TEST REPETITION :	yes	no

Table 2.1 : Parameters in Preference Testing

A good reference signal should have the following properties :

a) The reference signal should be variable in its quality between hifi quality and a not defined worst value, so that for all possible test signals there are corresponding isopreferent reference signals which are sufficiently inside the total quality range.

b) For accurate test results the reference signal should be similar to the anticipated test signals because the reliability of judgements on speech quality is influenced by the ease of comparisons between the test signal and the reference signal.

c) The reference signal and its generation should be exactly defined and allow for its simple and reliable reproduction in any laboratory.

d) The quality of the reference signal should be easily measurable.

e) It should be easily interpretable by engineers.

This list of requested properties makes idealized systems the most promising candidates for the derivation of a reference signal, as normal live systems cannot be expected to offer system conditions as closely defined and variable as necessary. Item b) has been stated although there is no conclusive theoretical way to define the variations and distortions of possible test signals which can be described in terms of a particular reference signal.

In order to get a simple measure for the degradation of a speech signal as requested in b) and c) a reference signal r(t) can be defined as a hifi speech signal s(t) plus a certain amount k of any distortion signal d(t). This basic assumption may be expressed by the simple equation

$$r(t) = s(t) + k.\,d(t)$$

The generation of this function r(t) is shown in Fig. 2.3

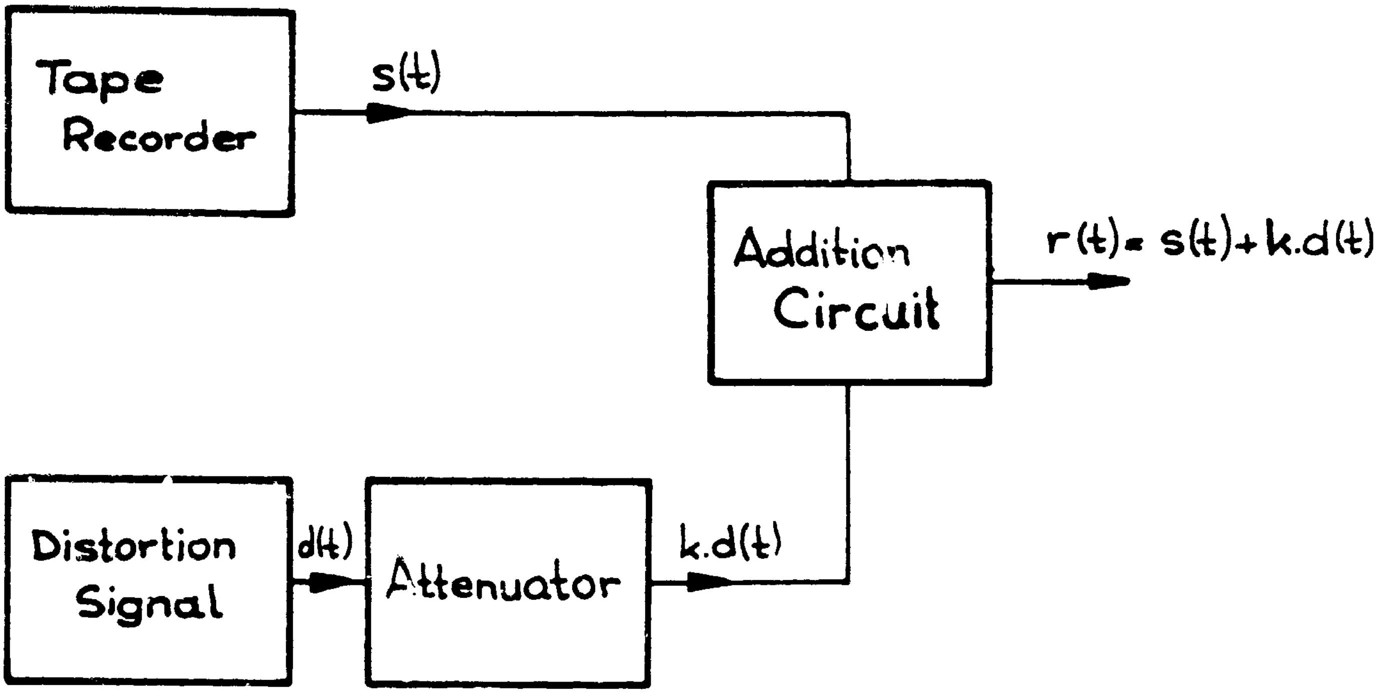


Fig. 2.3

Variations of the factor k yields a variable degradation of the hifi signal s(t) and consequently a variation of the speech quality of r(t) which can be easily expressed in terms of its S/N ratio. s(t) may be a real human speaker or only a hifi recording of such a speaker. Now the degradation signal d(t) has to be chosen, i.e. the question for a suitable reference signal has been changed to the question for only a suitable degradation signal.

Among all possible degradation signals those on the basis of random noise seem to fit best the desired properties a) - e).

Noise has several advantages including that of being physically measurable. White noise can be easily shaped by a suitable weighting curve to get a better approximation to average speech spectra. During our studies we have utilized different kinds of weighting networks :

u) A-noise : white noise, the spectrum of which has been shaped by an A-weighting network.

v) LP-noise : lowpass noise spectrum with a flat envelope up to about 500 cps and decay at a rate of 9 dB per octave above that frequency.

w) PINK-noise : noise with a reduction of the higher frequencies by 3 dB per octave.

Fig. 2.4 shows the three noise spectra

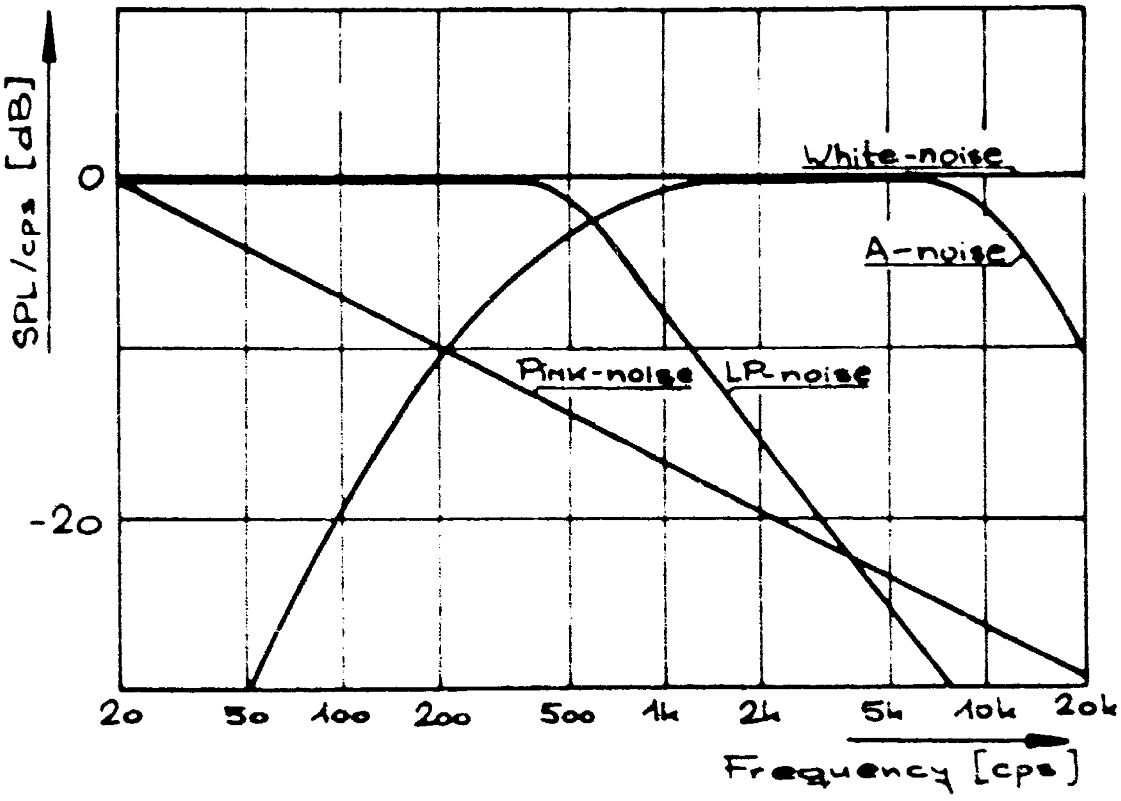


Fig. 2.4

From a technical point of view A-noise should be preferred. It is bandlimited on both sides and therefore avoids the problem of overloading the transducer with energy outside the normal hearing range. An additional advantage of the A-curve is its standardization by I. S. O for acoustic measurements. A filter with such a response can therefore be assumed to be easily available in acoustic laboratories. If the listeners are given a choice between the different types of noises, they seem to favor the pink noise, because it contains less energy at high frequencies. Comparative tests did not reveal any significant differences which would make the decision for one of the three types of noises easier. Results derived with one type of reference signal can be compared to results with another reference by merely adding a constant which compensates for the different spectral shapes of the degradation noises.

The actual generation of the reference signal is shown in Fig. 2.5. This reference signal r(t) will e called in the following "additive reference".

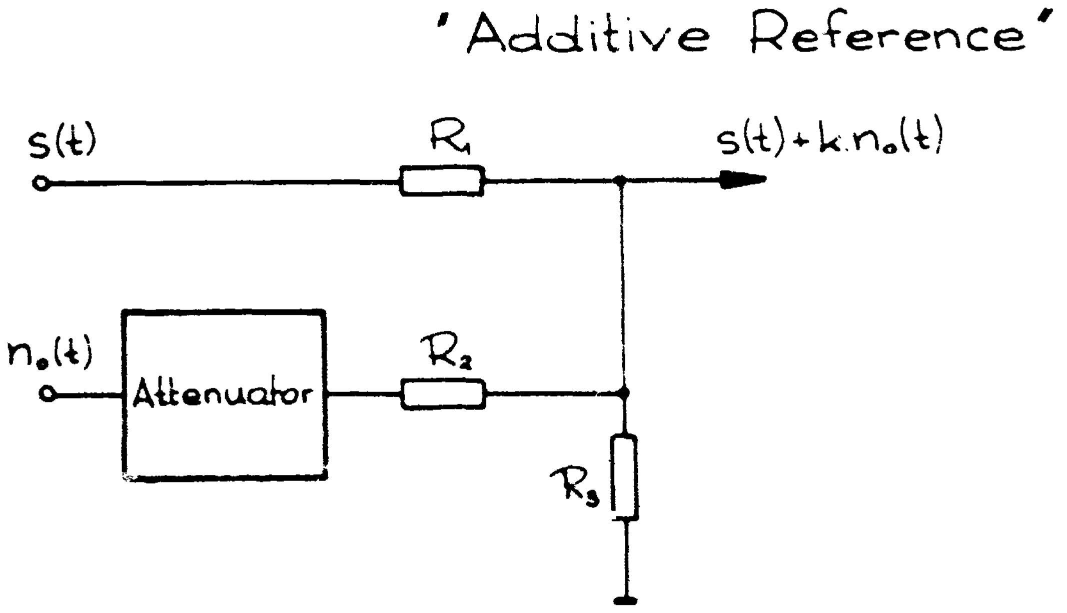


Fig. 2.5

The addition of the noise $n_o(t)$ to the hifi signal s(t) is probably the simplest way to generate a reference signal but this reference signal does not have all desired properties. Experience has shown that it does not comply with the properties listed under b) and d). It has been requested under b) that reference signal and test signal should be similar. The additive reference will only in a few practical cases satisfy this requirement. After becoming familiar with this reference signal most of the listeners are able to separate its two parts, i.e. they are aware that they hear hifi speech and simultaneously noise. The perception of this effect is enhanced by the fact that the noise degradation signal is always present. This may lead to difficulties especially when single isolated words instead of continuous texts are used as test material.

Another problem for the additive reference is listed under d). The quality of this reference signal is defined in terms of its S/N ratio which can be written as

level of the speech signal - level of the distortion signal.

While the noise level is measurable up to an accuracy of about 0.2 dB it is difficult to get a comprehensive definition of the speech level. It has been decided to circumvent this problem for the moment because the determination of the speech level with the desired accuracy turned out to be more difficult than anticipated. All our speech recordings carry therefore also a preceding pilot tone which allows to play the tapes always at the same level. The problem of the measurement of absolute speech levels could thus be postponed.

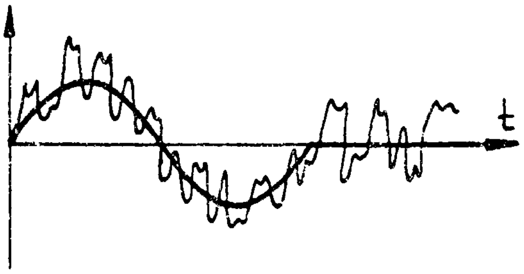
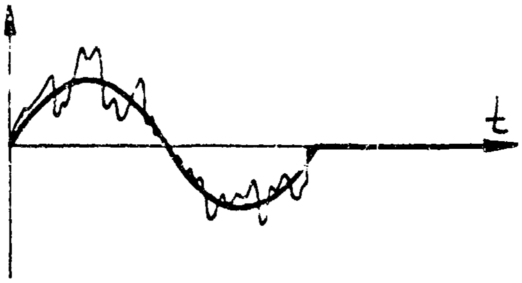
Considerations of the two problems of dissimilarity and of exact speech level measurements, mentioned above, recommend a search for other types of reference signals.

A promising distortion signal was found by multiplying the hifi speech signal with random noise.

$$d(t) = s(t) \ . \ n_o'(t)$$

The corresponding reference signal $r'(t) = s(t) + k.d(t) = s(t) . [1 + k.n_o'(t)]$ will be referred to as "Multiplicative Reference".

Fig. 2.6 demonstrates the main differences between the two reference signals. In contrast to the additive degradation signal, the multiplicative degradation signal is not present during speech pauses, thus it cannot be separated from the hifi speech signal by the listeners. Additionally, it overcomes also the second problem of the additive reference, as it does not require exact speech level measurements.

REFERENCE SIGNALS	
ADDITIVE REFERENCE	MULTIPLICATIVE REFERENCE
$s(t) + k.n_o(t) \quad 0 \leq k \leq 1$  	$s(t).[1+k.n_o'(t)] \quad 0 \leq k \leq 1$ 
$s(t) + k.n_o(t)$	$s(t) + k.s(t).n_o'(t)$
$S/N = 20 \log \frac{S}{k.N_o}$	$S/N = 20 \log \frac{\cancel{S}}{k.\cancel{S}.N_o'}$
$S/N = 20 \log 1/k$ – LEVEL OF NOISE SIGNAL $n_o(t)$ + LEVEL OF SPEECH SIGNAL $s(t)$	$S/N = 20 \log 1/k$ – LEVEL OF NOISE SIGNAL $n_o'(t)$

Here the speech level has not to be determined with high accuracy. The operation $s(t) \cdot n'_o(t)$ causes the distortion level to change in just the same way as the speech level. Therefore the S/N ratio of the multiplicative reference is independent from the speech level.

The better fit of the multiplicative reference to the list of required general properties than that of the additive reference reflects itself in the test results. On the average there are smaller standard deviations of the test results when preference tests are performed with multiplicative noise mainly because of the greater similarity of the test and reference signals.

The generation of the multiplicative reference signal $r'(t)$ is shown by the blockdiagram of Fig. 2.7. The multiplication of speech $s(t)$ and noise $n'_o(t)$ is done by a Hall-Multiplier,

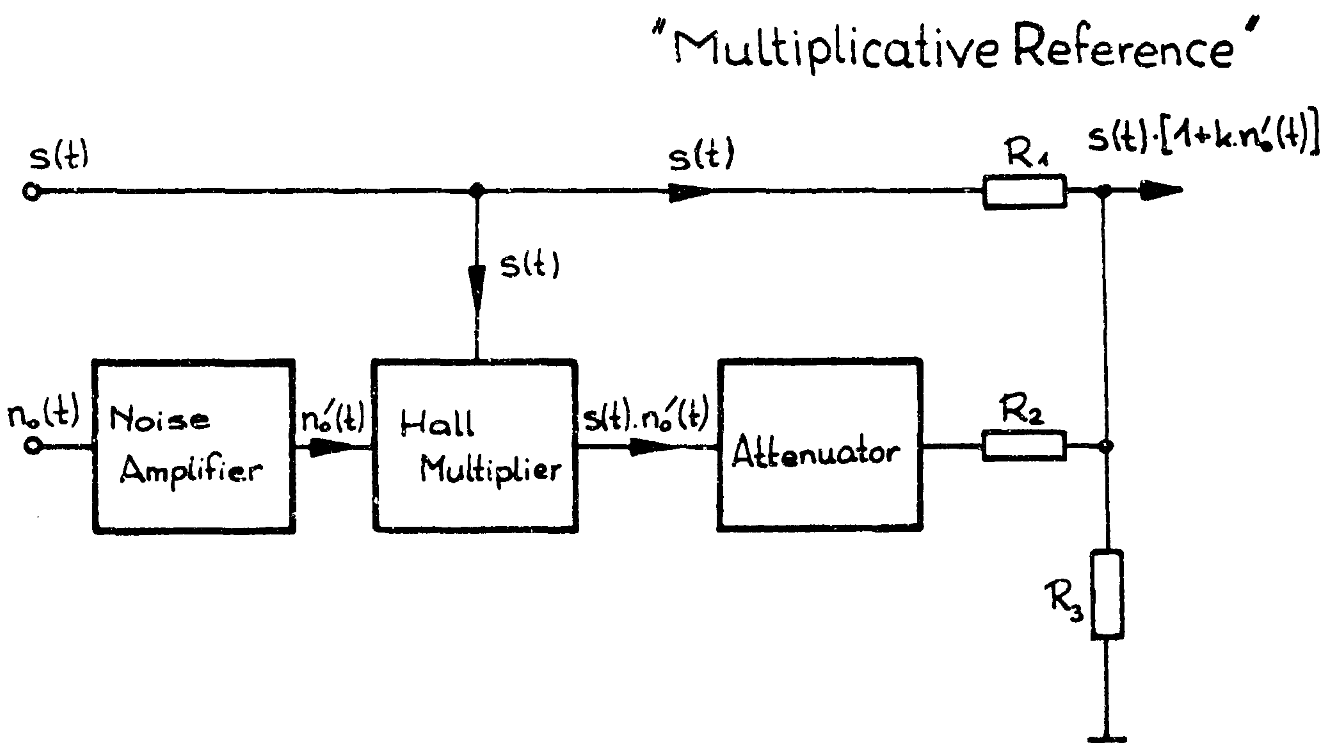


Fig. 2.7

a blockdiagram of which is shown in Fig. 2.8.

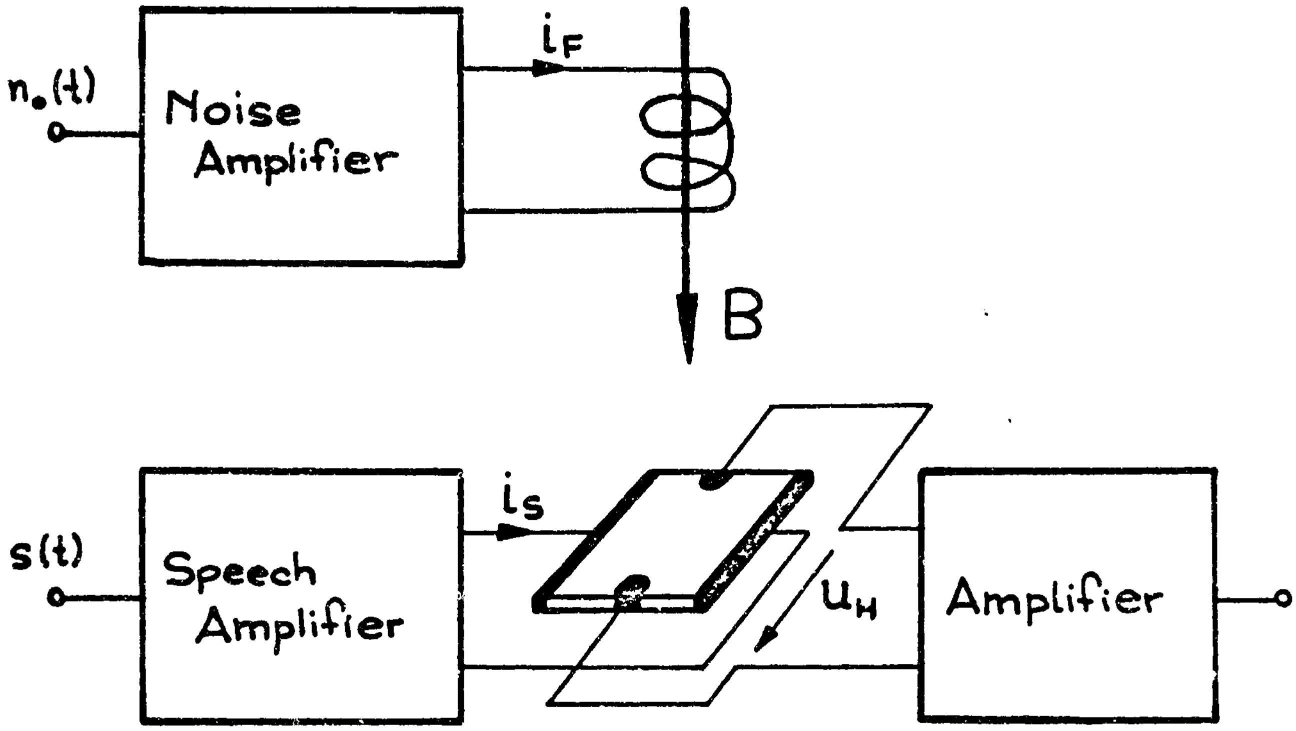


Fig. 2.8

The output voltage of this multiplier can be expressed by the equation

$$u_H = R_H \cdot i_s \cdot B + k_1 \cdot i_s + k_2 \cdot B$$

k_1, k_2 .... constants

R_H ...... Hall-constant

Making the control current i_s proportional to the speech signal s(t) and the magnetic field B caused by the field current i_F proportional to the noise signal n'_o (t), one gets the output

$$u_H = K \cdot s(t) \cdot n'_o(t) + K_1 \cdot s(t) + K_2 \cdot n'_o(t)$$

K, K_1, K_2 .... constants

The terms with K_1 and K_2 are deviations from the ideal product $s(t) \cdot n_o'(t)$. They are caused by non-ideal properties the so called "zero components" of the Hall-multiplier. A compensation of these terms by some special arrangements of the circuit is possible. Finally the resulting accuracy of the Hall multiplier is about 30 dB relative to the optimum output signal. A complete circuit diagram of the generation of the multiplicative reference is given in the Appendix.

The Hall-multiplier poses a second problem besides its zero components which were mentioned above. The magnetic field B is proportional to the field current i_F which has to flow through the fieldcoil with its high inductivity. The field current i_F now should have a spectrum according to that of the noise signal and because of the high inductive load, there are difficulties with the high frequency components of $n_o'(t)$. A power amplifier is necessary as a current source and additionally the spectrum of the noise input signal has to be pre-emphasized at high frequencies. The spectrum of the field current i_F and with it of the noise signal $n_o'(t)$ is given in Fig. 2.9.

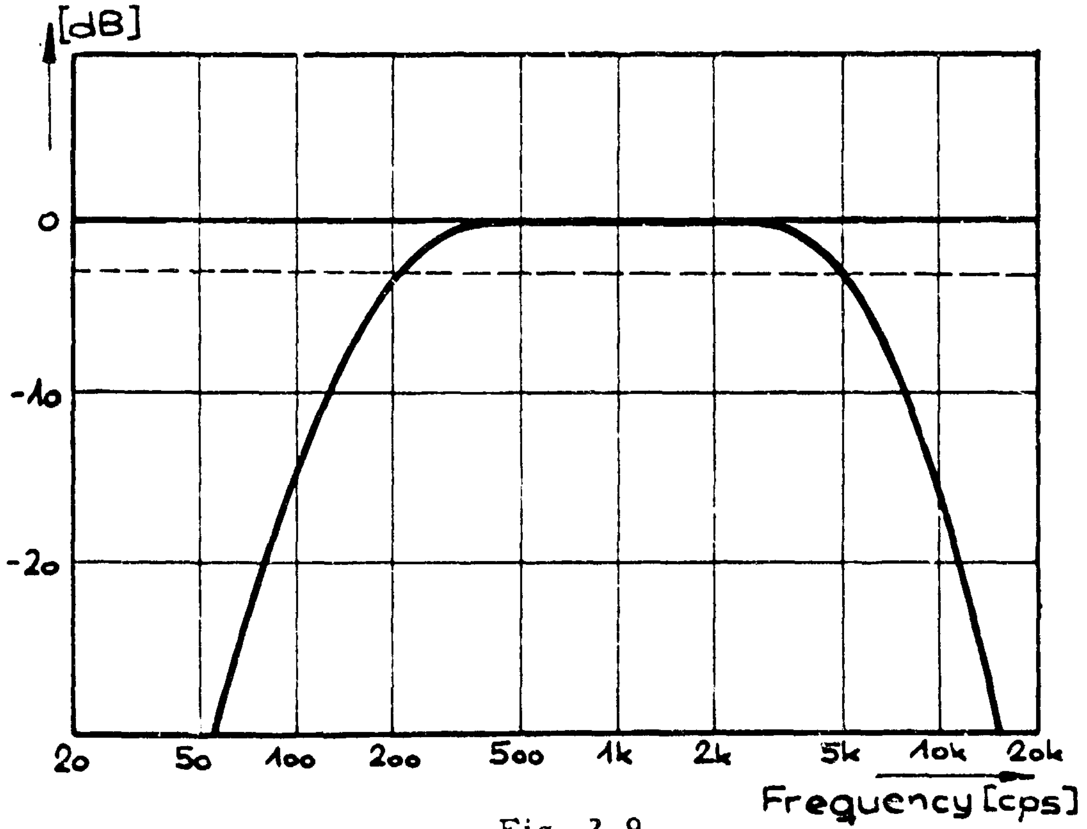


Fig. 2.9

Although, in most of our experiments we have generated the multiplicative reference by a Hall-multiplier, we were not too well satisfied with the performance of this electronic device.

It was therefore tried to find another more suitable form of generating the product of two signals. The new idea was to interpret the product of two suitable functions as a controlled switching process because a switch can be more easily and accurately implemented than an analog multiplier. The desired reference signal should have a noisy character. This could be achieved, e.g. by random interruptions or polarity inversions of the hifi speech signal. It was decided to utilize the second method of random inversion in periodic intervals. The pulse chain with random character for control of the inverter switch is derived from the output of a noise generator by sampling the noise signal periodically with a certain clock frequency.

The properties of the pulse chain can be specified by the clock frequency used and the probability that it will actuate the inversion switch in the sampling points. This probability was fixed to be 50 %. The not yet determined value of the clock frequency is chosen, so that the intelligibility of the product signal $s(t) . n_o (t)$ is as low as possible.

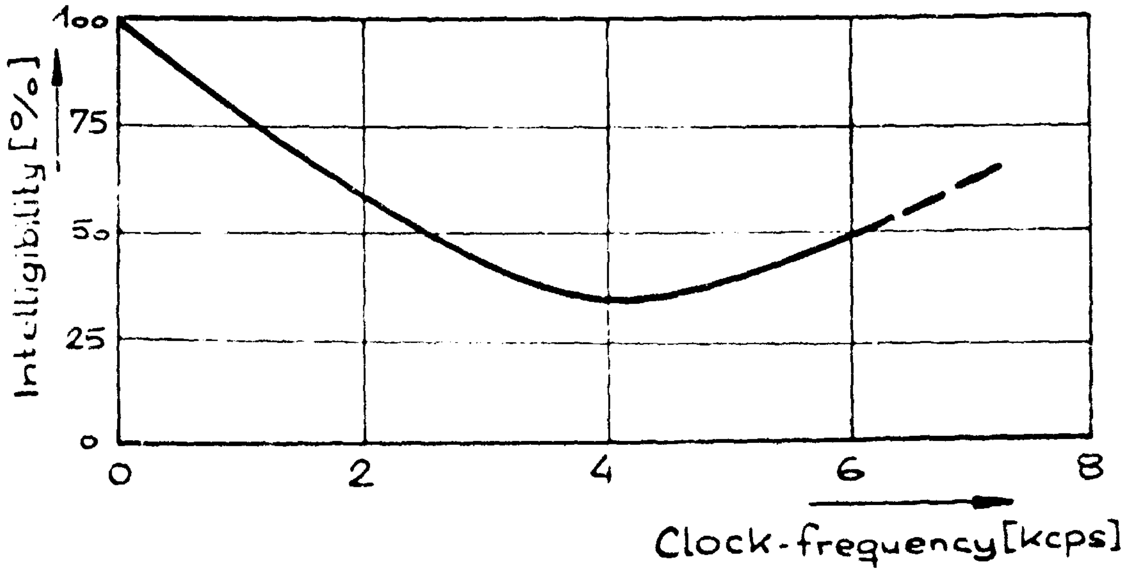


Fig. 2.10

The relation between intelligibility and clock frequency is shown in Fig. 2.10 and the minimum value yields a corresponding frequency of about 4 kcps.

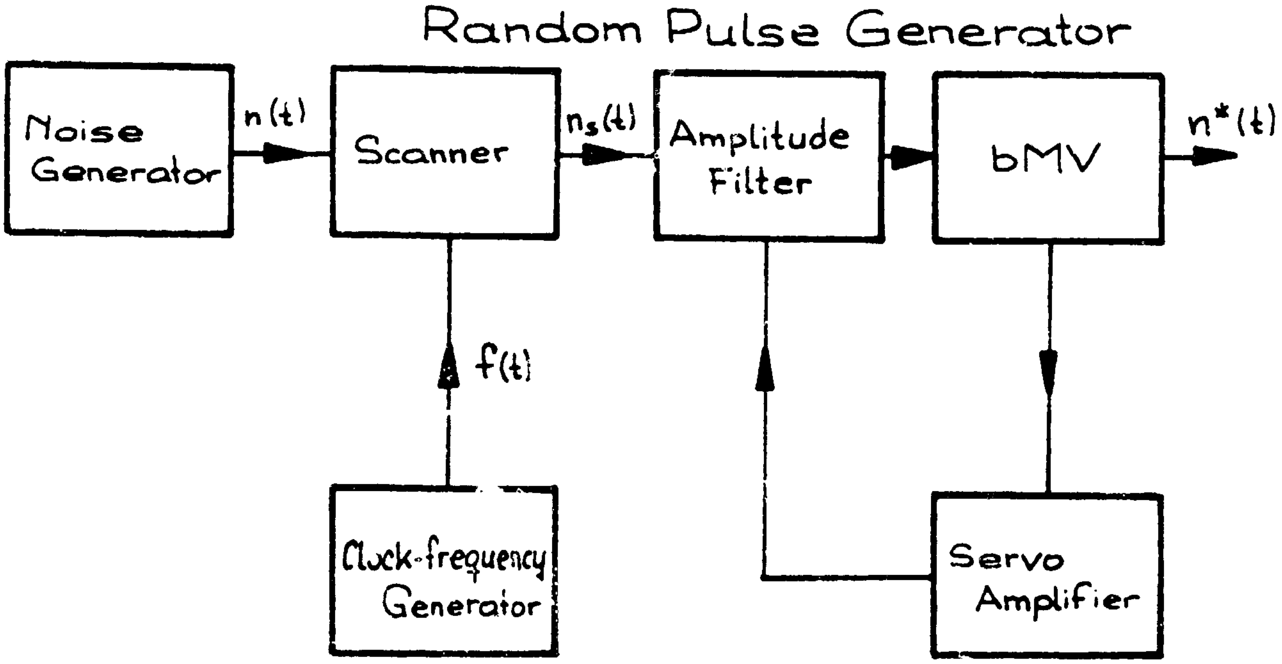


Fig. 2.11

The generation of the switching function $n^*(t)$ is shown in the block diagram of Fig. 2.11 and in the time table of Fig. 2.13. After sampling the noise signal n(t) by a scanner, the resulting pulses $n_s(t)$ are filtered by an amplitude filter. The remaining pulses control a bistable multivibrator bMV, which generates the switching function $n^*(t)$. A control loop including a servo amplifier is provided to ensure the 50 % probability in the switching points of $n^*(t)$.

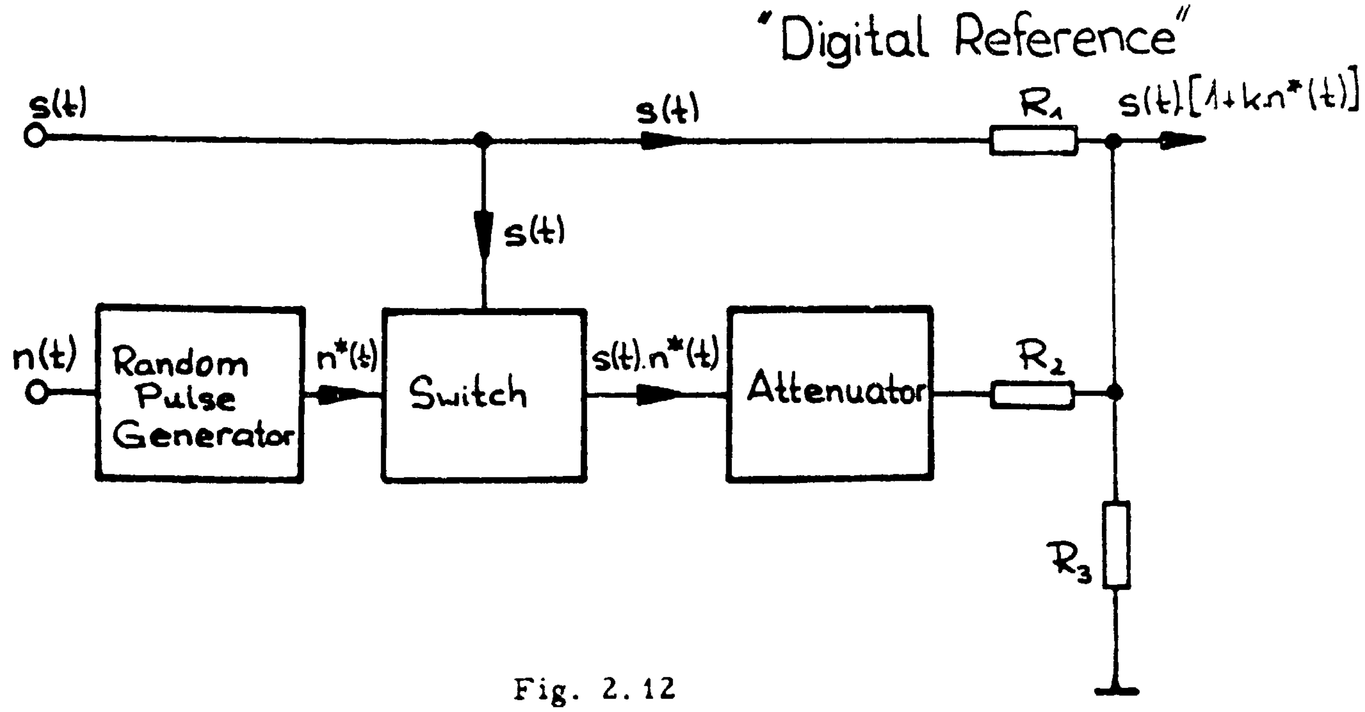


Fig. 2.12

The generation of the "digital reference" signal

$$r\ (t) = s(t) + k\ .\ s(t)\ .\ n^{*}(t)$$

which will be referred to as "Digital Reference" is shown in Fig. 2.12.

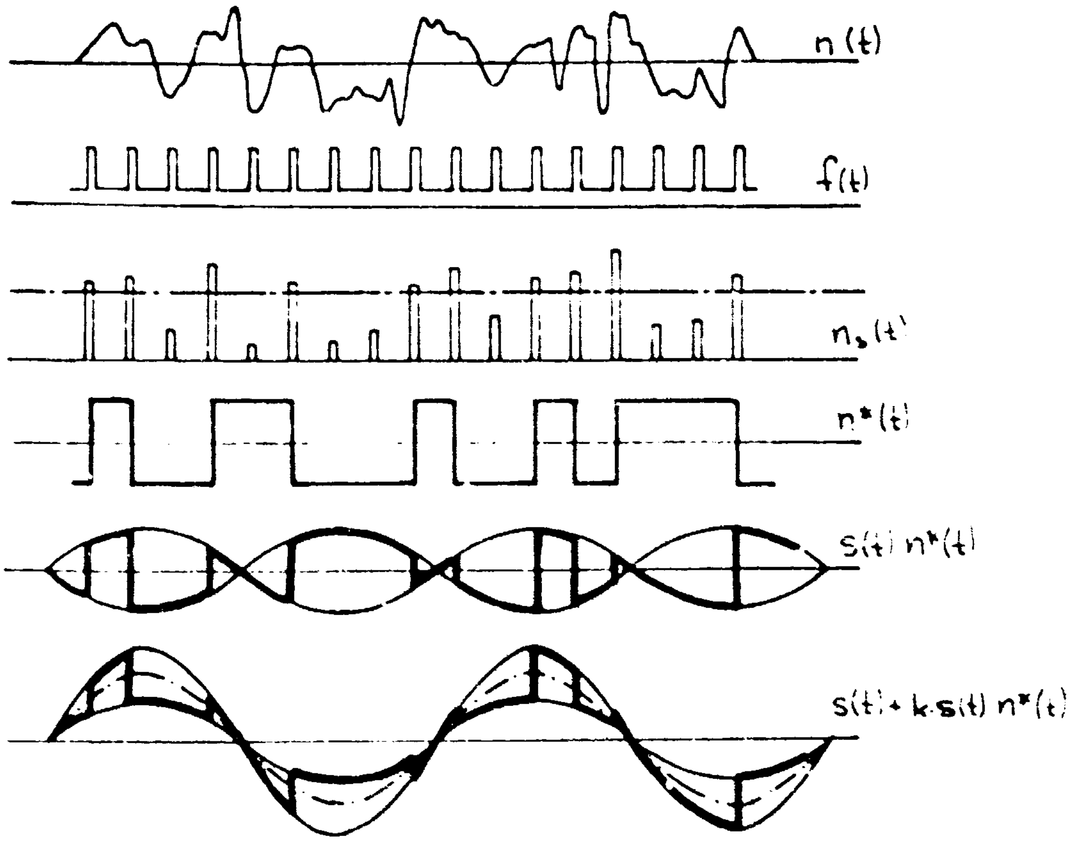


Fig. 2.13

The digital reference sounds similar to multiplicative reference. As its electronic implementation causes less problems than the Hall multiplier, it seems to be superior to the latter. More experience with this digital reference is still needed before we can make more definite recommendations. The Appendix contains detailed circuit diagrams for the implementation of this concept.

2.22 Test Signals

Speech test signals are output signals of any natural or artificial speech transmission, reproduction or composition system, the speech quality of which is to be evaluated. For our purpose of studying the usefulness of a method for preference testing its performance with a large variety of test signals should be evaluated. This is necessary in order to prove the reliability and justification of a procedure which uses only one kind of reference signal for comparison with the numerous possible variations of the properties of a speech signal.

The used speech test material consists of a continuous text and not of single words or syllables. The test signal is presented to the listeners with optimum loudness and is compared with a variably degraded reference signal. The determination of the optimum loudness of a special speech signal is discussed in chapter 3.13. For our preference tests we used a set of test signals produced by three different kinds of speech systems as shown in Table 1.1 :

a) LIVE SYSTEMS : Natural systems which are in a normal use as speech transmission or processing system.

a1) Telephone : Real local telephone circuit (Tel) using a transmission loop from one location over a dialled connection to the PBX and back to the same location.

a2) Vocoder : Channel vocoder from the IBM Laboratory Vienna /5/ in a special setting:

Fundamental frequency : normal (VON)
110 cps (VO1)
200 cps (VO2)

a3) Delta Modulation : Pulse modulation system in the following settings :

Sampling frequency : 7.2 kcps...(PD1)
20 kcps...(PD2)
43.2 kcps...(PD3)
60 kcps...(PD4)
120 kcps... (PD5)

b) IDEALIZED SYSTEMS : Artificial systems which produce an output signal consisting of a high fidelity recording of real speech, variably distorted by any form of additive or multiplicative noise.

b1) Additive Noise : Additive reference signal (ADD), hifi + k . noise.

b2) Multiplicative Noise : Multiplicative reference signal (MULT), hifi . (1 + k . noise); Hall multiplier.

b3) Digital Noise : Digital reference signal (DIG), hifi . (1 + k . noise); random pulse chain.

Because of the variable and adjustable degradation, first of all these speech signals are used as reference signals and therefore are described in detail in the previous chapter. Conducting preference tests between two of these speech signals, one of them may act as a variable reference signal; the other one may be held constant acting as the test signal. The results yield the important relations between the three reference signals which allow a first examination of the transitivity of the proposed method of preference testing.

c) SIMULATED SYSTEMS: Artificial system the properties of which are simulated by any conceivable distortion of speech signals, e.g. filtering, clipping, echo, crosstalk etc. of a hifi speech signal, or live systems with any additional artificial distortions.

c1) Lowpass : Filtered speech by a lowpass (LP) with a cut-off frequency of 1 kc. The rejected frequencies were attenuated at a rate of 40 dB per octave.

c2) Highpass : Filtered speech by a highpass (HP) with a cut -off frequency of 1 kc. The rejected frequencies were attenuated at a rate of 40 dB per octave.

c3) Live System with additive Distortion : Additional artificial distortions allow to increase the set of available test signals and to study the effects of superposed distortions.

2.23 Listening Group

A statement on the general acceptability of a particular system with regards to the public attitude towards a particular aspect of speech quality has to be based upon the judgements of a sufficient number of listeners. In order to prove the usefulness of the proposed preference method, it is necessary to describe the accuracy of such tests not only for a group of listeners at one time, but also for the single listener in repeated tests over a longer time interval. Training of a special listening group should ensure that the tests can be run under stable conditions.

At the beginning of a test session the listeners are informed about the purpose of the test and the testing procedure. In order to familiarize the test persons with the different types of test signals which will be encountered during the following test every new speech signal is presented for about two minutes before the actual test starts.

Three different groups of listeners have been utilized until now; all of them being male adults between 20 and 35 years of age :

a) A large group of untrained observers. Two times a year about 80 new students have to take laboratory exercises in our institute. Nearly none of them have ever been exposed to psycho-acoustic measurements.These listeners therefore are untrained and perform all the desired tests without test repetitions. The results of these groups should show the difference between a large group of untrained and a small group of trained listeners and hopefully also indicate the "optimum" number of listeners with regards to a compromise between financial expenses and "statistical" significance of subjective tests.

b) A small group of about 20 trained persons. This testing group was used for the collection of most of the data contained in this report. All listeners of this group were examined for normal hearing. They meet the requirements on auditory acuity in the American Standard on Measurement of Monosyllabic Word Intelligibility /3/. We have found that these measurements could be replaced for our purposes by the correct response to an intelligibility test with monosyllabic words. Naturally we have utilized German word lists for our students /4/. With this listening group not only the preference measurements were conducted but they helped also to study side effects such as training and learning, fatigue, reproducibility etc. From the first 20 students about a dozen,has carried on until now, the others had to be replaced due to lack of time or interest. Our test room facilities accommodate a maximum of 10 listeners at one time. We therefore had to form two groups of 10 listaners each which helped also to satisfy the different working time schedules of the listeners. Both groups were exposed to practically the same test material. The number of listeners at one test session was about 8. The duration of one session consisting of about 6 test runs, was two hours in the average.

c) A very small "special purpose" testing crew. This group consists of ourselves and staff members of our institute. Besides gaining the necessary possibility to understand from personal experience also the listeners' position, we attacked side problems such as optimum loudness, check-on-order effect, difference-limens and "critical ranges".

The question when a single listener or a group may be qualified as "trained" for preference tests is not easily answerable and will be discussed in Sec. 3.

2.3 Test Set-up

2.31 Test Environment

A room for psychological measurements should be reasonably free of inside and extraneous noise. Therefore it is practical to provide different rooms for the listeners and for the operator with his equipment.

As only headphones for the presentation of the acoustic stimuli were employed there is no necessity for the installation of an anechoic chamber. Furthermore the utilized KOSS-PRO-4 headsets have soft earcushions for additional protection against ambient noise. A test room with studio character having a small reverberation is therefore adequate. A quiet groundfloor laboratory room has been modified for our purposes. The listener room of approximately 23 x 9 x 10 ft in size has the walls covered with perforated boxlike aluminium sheets. A lightweight blanket of glasswool is laid behind them to provide sound absorption. The ceiling consists of suspended broadband-absorber units of a styropor-like foam material. For a reduction of the noise level inside the room caused by extraneous disturbances the entrance has been improved by installation of a second sound insulating door. A measurement of the reverberation time yielded an average value of 0,22 s for the empty room and a value of 0,2o s for the room when occupied by the listeners. A photo in the Appendix shows the interior of the room.

2.32 Signaling System

Psychological testing is very time-consuming and it is desirable therefore to conduct the tests as efficiently as possible. In the listening room accomodations for 10 listeners have been installed. As the operator and the test equipment are located in a separated adjacent room, besides all necessary equipment and connections for the presentation of speech signals, also a signaling system had to be provided. It has been designed and built with an aim towards automatic operation and for minimizing the possibilities for mutual influence among the listeners. For storage and recording the test results are not only displayed on a lamp panel, but can also be printed by an automatically controlled teletypewriter. An inter-communication link allows for conversation between the listeners and operator.

Fig. 2.14 shows a block diagram of the signaling system.

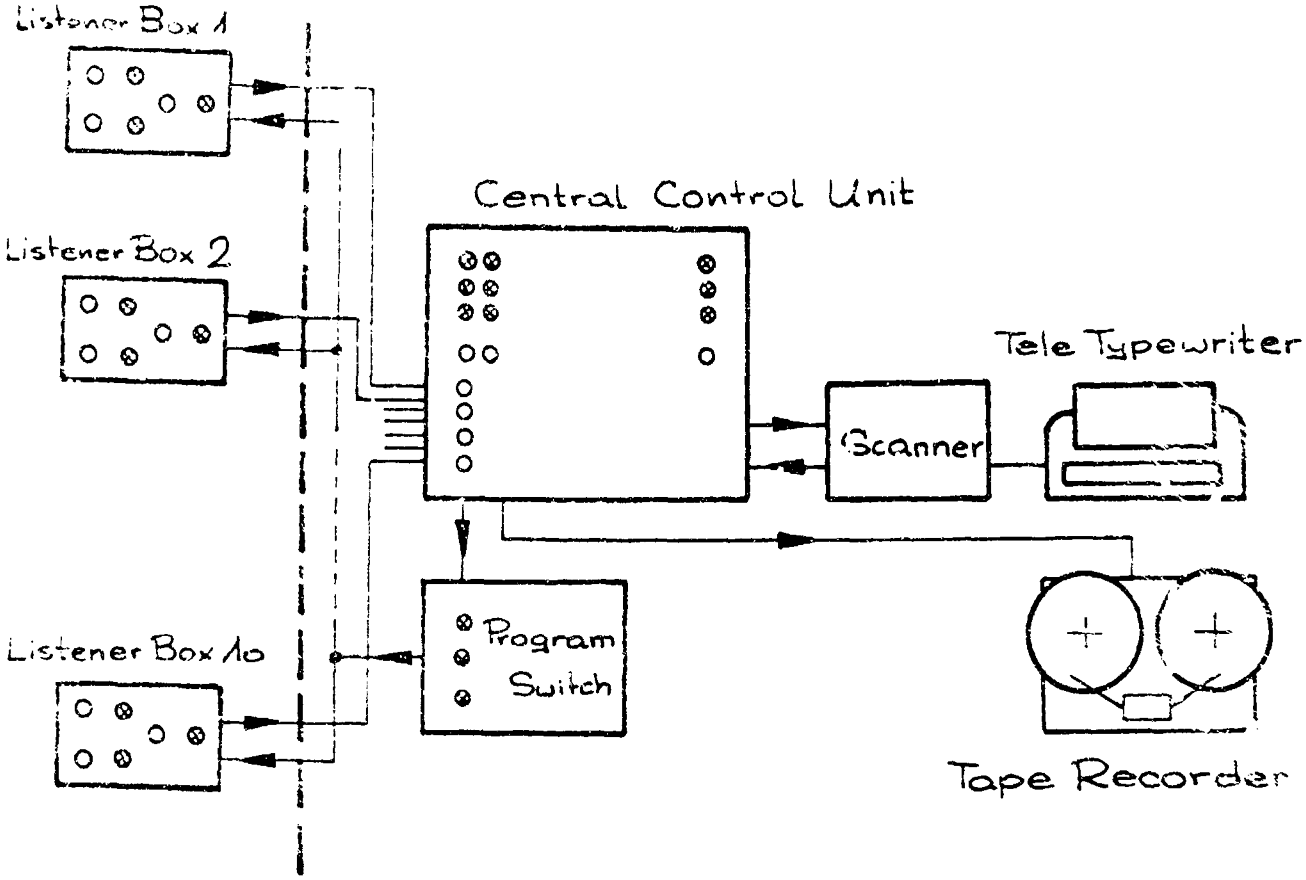


Fig. 2.14

In the listener room there are ten terminals with listener sets consisting of earphones and the "listener boxes". On each of the boxes three small lamps "A", "B", and "READY" are mounted which belong to three corresponding push buttons "A", "B", and "STOP". During the test run the signal lamps A and B are controlled by the program switch and indicate the corresponding speech signals as they are presented. After presentation of a repeated signal pair, each listener indicates his decision by pushing the corresponding button A or B. Now the "READY" lamp serves as indicator for the listener that he has taken his decision. After a wrong decision or for some other reasons each listener can stop the test run by pushing the "STOP" button.

In the operator room a light display and control unit stores the listener's decisions in a bank of relays. The visual indication is done by corresponding lamps. After all decisions have been received a scanner automatically reads the results into an electric teletypewriter. Then all lamps are reset and a starter impulse to the program switch starts a new test run.

A block diagram of the signaling system between parts of the test equipment and the principal connections between listener and operator room can be found in the Appendix.

2.33 Test Equipment

A block diagram of the test set-up for conducting pair-comparison tests is shown in Fig. 2.15. This set-up is utilized for our experiments. Two channels are provided for two different audio signals, for the present study mostly speech signals. Controlled by one noise generator two separate distortion signals can be generated which allow for independent degradation of the signals in the two main channels. A program switch controls the presentation of the acoustic stimuli to the listeners who can then make the desired observations with regard to any special property of the samples, e.g. preference, loudness etc.

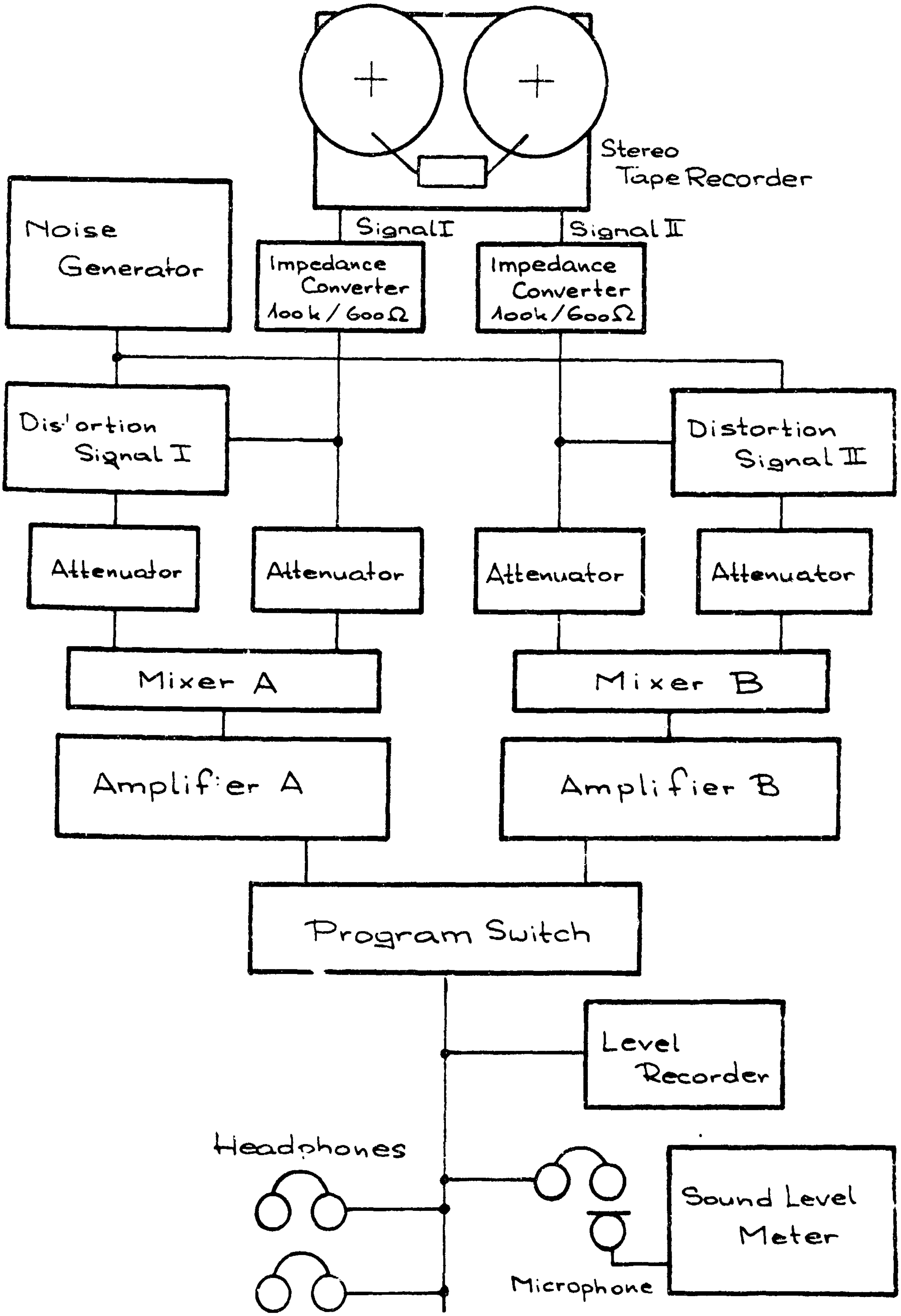

Stereo
Tape Recorder
Signal I
Signal II
Noise Generator
Impedance Converter 100k/600Ω
Impedance Converter 100k/600Ω
Dis'ortion Signal I
Distortion Signal II
Attenuator
Attenuator
Attenuator
Attenuator
Mixer A
Mixer B
Amplifier A
Amplifier B
Program Switch
Level Recorder
Headphones
Sound Level Meter
Microphone

Fig. 2.15

2.34 Hifi Speech Signal

In contrast to the experiments reported by MUNSON and KARLIN we did not use a real speaker for presentation of speech signals to the listeners. A hifi tape recording avoids all the difficulties which arise when running subjective tests with a real speaker. Among others the advantages of a high fidelity tape recording are the unlimited reproducibility of the speech signal with invariable articulation and reproducible loudness level so that the tests may be repeated with identical signals as often as desired. The utilization of a real speaker would create another serious problem. Obviously, his utterances could not be presented over headphones. The whole mode of presentation would have to be changed. In contrast to these problems with a real speaker, the main disadvantage of tape recordings is their limited quality. As we are now only interested in the evaluation of test signals which have "telephone quality", our hifi recordings are still by far superior in quality. Therefore changes in the quality ratings we find are not to be expected, if a real speaker instead of tape recordings were used during the test.

The speech material used in our tests is taken from a master tape which has been prepared previously at the IBM Laboratory Vienna. A professional radiospeaker has read public news under studio conditions. The recordings are made by means of a dynamic AKG microphone type D 20 B and an Ampex 351 tape recorder. The frequency range of the recordings is better than ± 3 dB from 50 to 15.000 cps. The ambient noise conditions during the recordings yielded a S/N ratio on the master tape of about 50 dB.

On the master tape there is additionally a 1 000 cps pilot tone as reference for the purpose of level measurements. Besides the continuous text, 400 monosyllabic German words have been recorded under the same conditions to be used for intelligibility testing.

The tapes actually used for testing are prepared in the following manner. A re-recording of the master tape from AMPEX to the first track of the test tape has been made using a REVOX G 36 tape recorder. Connecting the output from the AMPEX with the input of the system to be evaluated, the system output signal then is recorded on the second track of the test tape preferably text-synchronous to the text on track 1. Each recording is preceded again by a pilot tone for convenient adjustment of the speech level. After recording the 400 single words having passed the system to be tested, the preparation of the test tape is finished. The content of the test tape is shown in Fig. 2. 16

track 1	1000 cps tone	hifi speech signal	200 words testquality
track 2	1000 cps tone	test speech signal	200 words testquality

Fig. 2. 16

Accurate level measurements of speech signals are a well known problem. We could circumvent this problem for the work reported here. As we wanted to study mainly the test procedure and the reproducibility of the listeners' decisions there was no need for the absolute speech level measurements which are mandatory for absolute quality ratings with the additive reference. It was only necessary to keep the speech level as constant as possible during the recording sessions and to take care for reproducibility during the reproduction of speech material. As already mentioned a pilot tone was used for initial level adjustment and a graphic level recorder for continuous monitoring of the speech level. Obviously, difficulties may arise, when the original speech material, i.e. the master tape is changed. Instead of establishing electrical reference conditions it was decided to refer always to the acoustical input to the ears of the listeners.

When the speech samples are presented to the listeners via earphones the sound pressure level of the speech signal is measured objectively with an artificial ear. Considering the well-known difficulties with the standardized artificial ear for higher frequency measurements and its proper acoustic coupling to headphones with earcushions like our KOSS PRO-4, a simplified construction was used which gives at least reproducible results. Fig. 2.17 shows a sketch of our arrangement of a wooden plate with an inserted condenser microphone. The earphone is pressed to the plate with about 1 000 g. The air volume remaining between the plate and the membranes is about 12 cm^3.

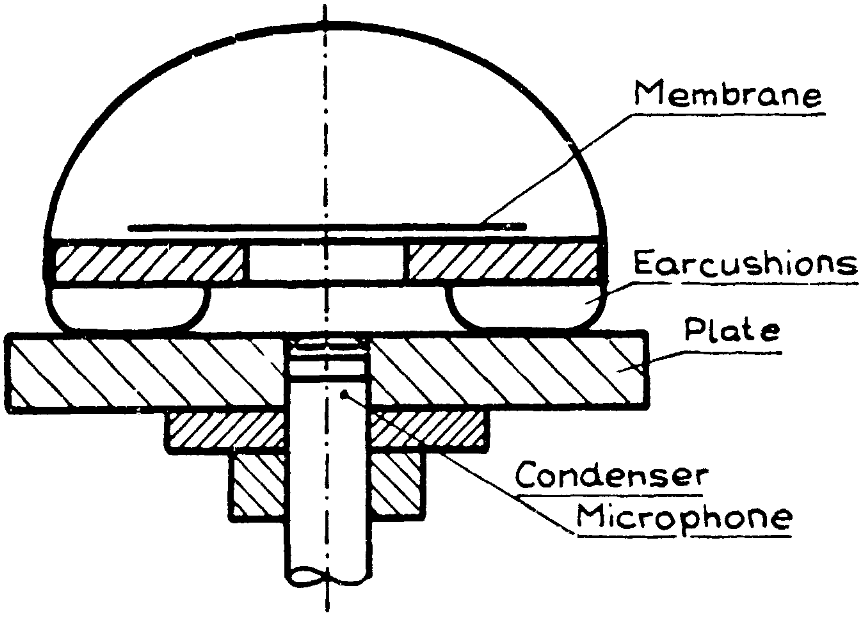


Fig. 2.17

The chosen procedure for speech level measurements tries to follow a practice for the measurement of certain impulsive noises. There the level is defined as the arithmetic average of the maximum values of the A-weighted sound levels. This average is determined by considering only those maximum values which are within 10 dB of the highest occurring value. The average has to be taken over a suitable period which may coincide in our case with the duration of the speech sample to be measured.

A subjective method for the measurement of speech level especially with respect to the actually used S/N ratio will be discussed in Sec. 3.13.

2.4 Analysis of Test Data

In preference tests listeners have to decide whether they prefer the test signal A to reference signal B or vice versa. As listeners are forced to decide either for signal A or for signal B the test data form a complete sample space with the reference signal given by its S/N ratio figures as a parameter. Data collected in several separate tests can be treated together for evaluation as other statistical material.

The following considerations may form a basis for improvements and refinements of the test procedure and test evaluation as they are performed now. In the next subsections several theorems and relations known from mathematical statistics will be used without giving any proofs. The interested reader should refer to the pertinent literature. Our considerations shall only give some suggestions for handling the data obtained by preference tests.

Although the actual tests are run with groups of listeners, first of all it is useful to consider the decisions of an individual listener under test. In a second subsection groups of listeners will be considered and finally our present method will be described which is utilized for the processing of test data.

2.41 Basic Statistical Considerations

The idealized individual listener

For the following considerations we define the idealized individual listener as a person who decides in a manner that the relative frequency of preferring signal B to signal A converges to the corresponding probability for any S/N ratio of signal B. Further we assume that the probability of preferring signal B to signal A does not decrease when the S/N ratio of B increases. This fact is shown in Fig. 2.18 where p_x denotes the probability that signal B is preferred to signal A at a S/N ratio of x dB for signal B. The function p_x verse the S/N ratio x will be called "graph of preference" for abbreviation.

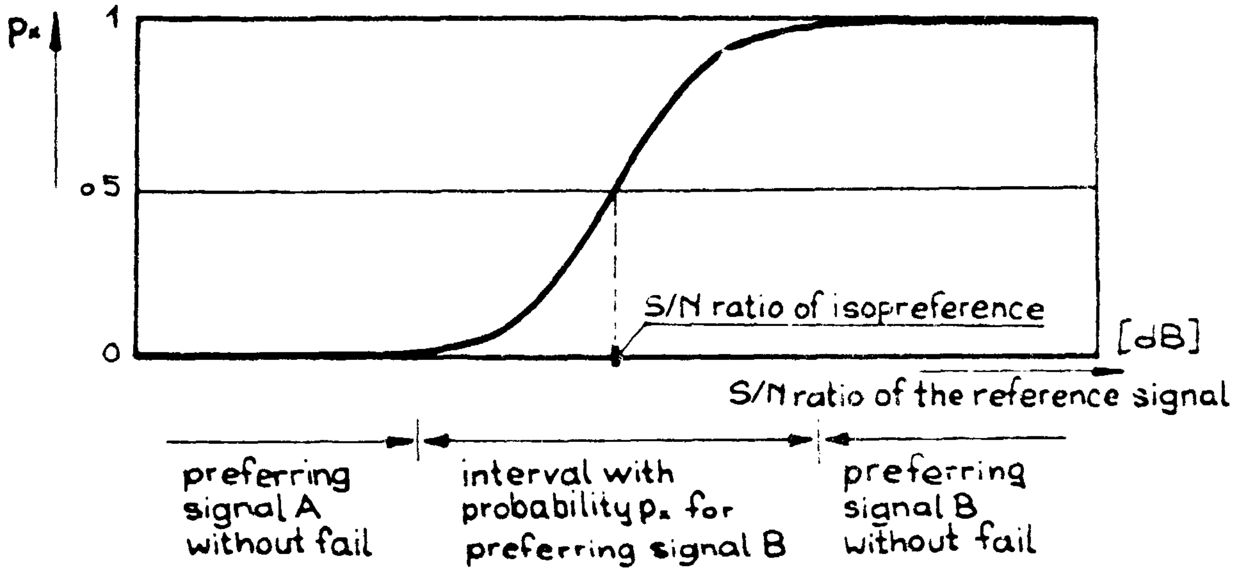


Fig. 2.18

Our assumption implies a stationary behaviour of the idealized listener, i.e. he should have a fixed opinion about the quality of the test signal. But this assumption does not pay regards to any effect of learning, training, accustoming, fatigue etc. which may occur when tests are performed over a long period of time. Our results show that a real individual listener is in good agreement with the ideal listener postulated above within a test period of several hours.

In Fig. 2.18 the abscissa of the graph of preference can be divided into three intervals : in the interval on the left hand side signal A is always preferred to signal B so that decisions within this interval will have a high amount of certainty. Similarly signal B will be preferred with a high amount of certainty when its S/N ratio is located on the right hand side of the abscissa. For S/N ratios of signal B within the middle part of the abscissa decisions of our listener will contain a random element, but in p_x of a large number of identical comparisons at a certain S/N ratio x signal B will be preferred to signal A. Within this middle part of the abscissa there also exists a special S/N ratio at which the individual listener will find both signals A and B "isopreferent", that is when the probability of preferring B to A is just equal to 50 %.

For the moment we are mainly interested in signals B which are located in that middle part of the abscissa mentioned above where dicisions of the idealized listener change between preferring signal B to signal A and vice versa with a frequency indicated by the graph of preference shown in Fig. 2.18. Suppose a test is run presenting n times the S/N ratio of x dB to our idealized individual listener under the same circumstances. This test then forms a Bernoulli test consisting in a succession of n Bernoulli trials which means each trial is performed under identical premises with probability p_x for preferring signal B and $(1 - p_x)$ for preferring signal A since A figures for the contradictory statistical event.

For n Bernoulli trials the number S_n of preferring signal B to signal A is a random variable which is binomially distributed /6/.

$$P(S_n = k) = \binom{n}{k} \cdot p_x^k \cdot (1 - p_x)^{n-k} \qquad (2.1)$$

This equation specifies the probability of S_n equal to k decisions preferring signal B at n Bernoulli trials with the probability p_x for preference. In this equation p_x figures as a parameter. For a special example the probability $P(S_n = k)$ is shown in Fig. 2.19.

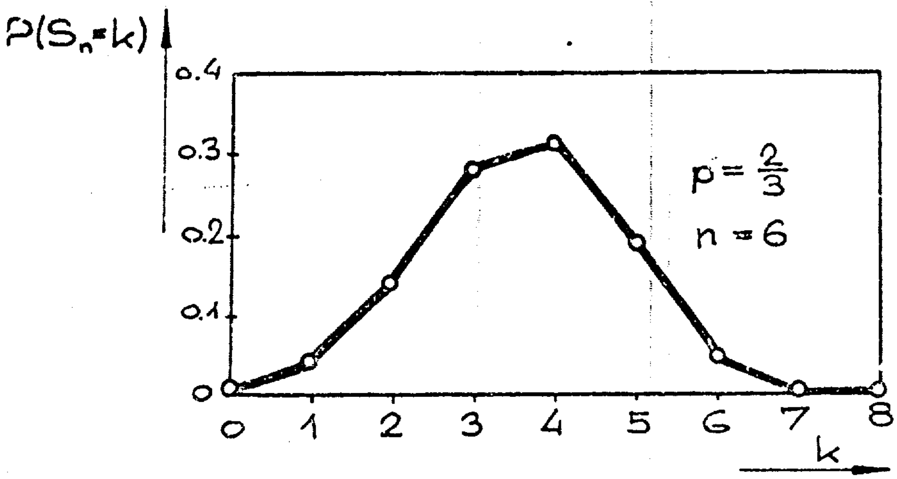


Fig. 2.19

As k goes from 0 to n $P(S_n = k)$ first increases monotonically reaching its greatest value for k = entier $(n + 1)p$ and then decreases monotonically. Further typical data for the distribution of $P(S_n = k)$ are the moments of the distribution. For the binomial distribution $P(S_n = k)$ the expectation or the first moment is given by

$$E(P(S_n = k)) = n \cdot p_x \tag{2.2}$$

and the standard deviation or the positive square root of the second moment with respect to the expectation is given by

$$\sigma(P(S_n = k)) = {}_{+}\sqrt{n \cdot p_x(1 - p_x)} \tag{2.3}$$

Both values are better considered in proportion to the n trials of which the assumed Bernoulli test consists. The expectation divided by the number of Bernoulli trials is equal to the probability p_x of preferring signal B. At the same time the standard deviation divided by the number of Bernoulli trials becomes inversely proportional to n which means that for decreasing the standard deviation by a factor c the number of Bernoulli trials must be increased by a factor c^2.

Now Bernoulli tests shall be utilized to evaluate the probability p_x of preferring signal B to signal A at the S/N ratio x dB. From the relative standard deviation of the binomial distribution

$$\frac{\sigma(P(S_n = k))}{n} = \frac{\sqrt{p_x(1 - p_x)}}{\sqrt{n}} \qquad (2.4$$

follows a constant value of $\frac{\sigma}{n}$ at different probabilities p_x the number of Bernoulli trials may be reduced as p_x approaches 0 or 1. Accordingly the number n of Bernoulli trials should be a maximum for a desired accuracy of p_x when p_x = o, 5.

Furtheron an estimate can be given for the reliability of results obtained in a test consisting of n Bernoulli trials by Laplace's limit theorem which holds for a large number of trials /7/.

$$P\left\{a \leq \frac{\frac{S_n}{n} - p_x}{\sqrt{\frac{p_x(1 - p_x)}{n}}} \leq b\right\} = \Phi(b) - \Phi(a) \qquad (2.5)$$

In this equation $\Phi(y)$ stands for the standard normal distribution function and a and b are preannounced limits. In this manner Equ. 2.5 yields the probability that the number S_n of decisions for B in n Bernoulli trials is sufficient to determine p_x within the limits indicated in this equation.

For determining a graph of preference for the individual listener as it is shown in Fig. 2.18 one should run Bernoulli tests with several different S/N ratios of signal B. Since we are mainly interested in the S/N ratio of isopreference, i.e. the S/N ratio where the graph of preference crosses 50 %, it is possible to shorten the test procedure by the following considerations. Several Bernoulli tests will be performed to find the two ranges of S/N ratios where the individual listeners prefer unambiguously A or B. One finds these ranges in Fig. 2.18 partly on the left side and partly on the right side of the abscissa. It is obvious that under these circumstances the listener will decide more or less without fail. As for these ranges of S/N ratios the probability of preferring signal B indicated by p_x is close to 0 or 1, results with small deviations can be obtained already by a small number of test runs.

Let us assume that a S/N ratio specifies the point of isopreference where for m Bernoulli trials, the individual listener votes m/2 times for signal B. By this we can find a distribution function Q and a density function q for the S/N ratio of isopreference with regard to m Bernoulli trials from the graph of preference of the individual listener (Fig. 2.20).

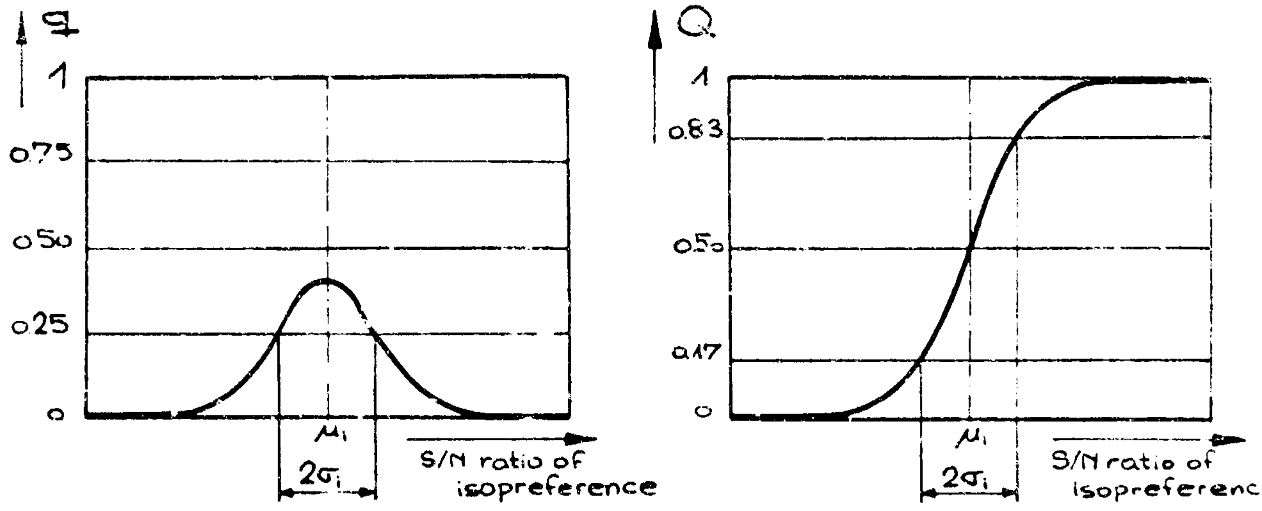


Fig. 2.20

A derivation of the distribution and density functions from the graph of preference shall not be given here, but it may be accepted that these functions can be replaced approximately by normal distributions with the parameters μ_i and σ_i. The subscript i denotes the individual listener. Obviously, the standard deviation σ_i of this distribution depends on the number of Bernoulli trials considered. For an infinite number of Bernoulli trials the distribution function degenerates into a step function and the standard deviation becomes zero. In this case the S/N ratio of isopreference for the individual listener will be fully determined.

By taking advantage of these considerations we can find the S/N ratio of isopreference for an individual listener when we perform several Bernoulli tests each consisting of m trials at S/N ratios where the decisions of the listeners under test are fairly unambiguous. As it is shown above the decisions of our listener will not vary very much at those S/N ratios and they will be therefore very certain. Thus we can get the left part and the right part of the distribution function of Fig. 2.20 without fail and can find now the middle part of it by interpolation. The S/N ratio where this distribution function crosses the 50 % ordinate will be a good estimate for the desired S/N ratio of isopreference for the individual listener.

For abbreviation we will speak further from a distribution of the S/N ratio of isopreference omitting to emphasize the number of Bernoulli trials necessary for it. By introducing the distribution of the S/N ratio of isopreference one may reduce the number of tests which are necessary for determining the S/N ratio of isopreference. This leads to a significant reduction of the necessary effort in preference testing. The exclusion of the transition region for the taking of sampling points will reduce the accuracy of the test results, but it will still be comparable to the accuracy limited by the technical facilities of the test set-up.

A group of listeners

The considerations concerning the individual listener shall be extended to a group of listeners. We take the group of listeners to be random samples from a population of listeners and their individual S/N ratios of isopreference as the statistical variable which is assumed to be normally distributed. The parameters oi this distribution are called μ_g for the mean and σ_g for the standard deviation. The subscript g refers to the group of listeners. The mean of this normal distribution indicates that 50 % of the listeners will have their S/N of isopreference lower than μ_gdB and therefore 50 % of the listeners will prefer the reference signal B to the test signal A when the reference is presented with a S/N ratio of μ_gdB. For the measurement of speech quality we are more interested in the S/N ratio of isopreference μ_g for the group of listeners than in the decisions of a single listener. The standard deviation $\sigma_g$of the normal distribution assumed is a measure for differences at the S/N ratios of isopreference for the individual listeners and may be of interest as far as the reliability of the test results is concerned.

The determination of the S/N ratio of isopreference for a group of listeners should start with isopreferent S/N ratios of the individual listeners. By plotting the percentage of those listeners whose S/N ratios of isopreference is lower than x dB versus the S/N ratio x. the experimental distribution function will be a staircase function as shown in Fig. 2.21

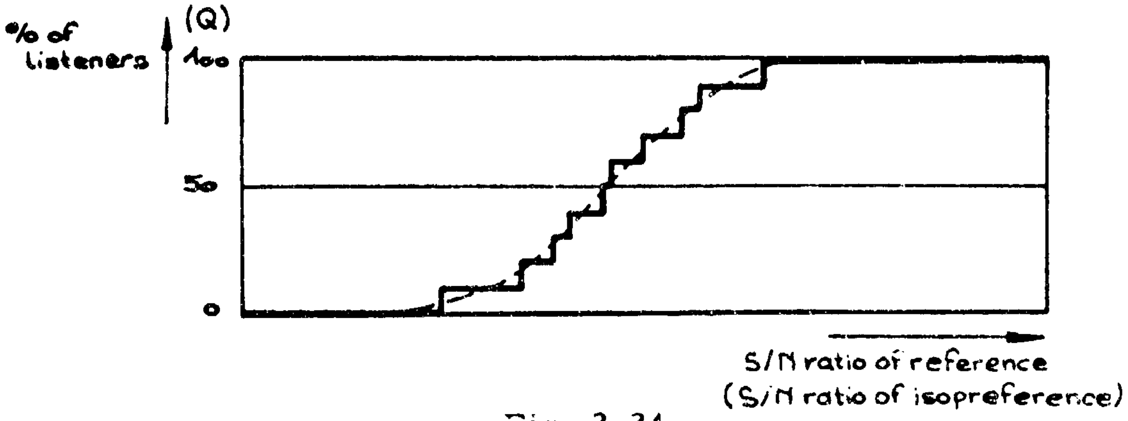


Fig. 2.21

This experimental distribution function approximates the assumed normal distribution function with the parameters μ_g and σ_g. This way is very cumbersome because at first all S/N ratios of isopreference for the individual listeners have to be calculated only then the desired S/N ratio of isopreference for the total group may be estimated. Therefore another method shall be described for the evaluation of the S/N ratio $\mu_g$of isopreference for the group which is not as "accurate" as the method mentioned above, but which is much easier to carry out and which still yields sufficient accuracy.

A simplified case shall be considered first : we assume graphs of preference for all individual listeners in form of simple step functions. Of course that is only a rough approximation to reality, yet it is very helpful for introducing the following method into the present concept. With this assumption it follows that at any S/N ratio of the reference each listener votes without fail either for signal B of for signal A respectively at any number of trials performed. Therefore one can take a test procedure in which each S/N ratio of the reference is just once presented to the group of listeners, and one then collects their decisions. This simple procedure yields here already the experimental distribution function of Fig. 2.21.

If one goes back to the real test conditions, graphs of preference for the individual listeners may look like Fig. 2.18. Running the same test procedure described just above we will have several listeners in the group voting not in accordance with their S/N ratios of isopreference. Plotting again the percentage of listeners preferring signal B versus the corresponding S/N ratio one finds that this empirical distribution function is not necessarily a monotonical increasing function as it should be. This is caused by the "fail" votes of listeners (Fig. 2.22). Still we may approximate this empirical distribution function by a normal

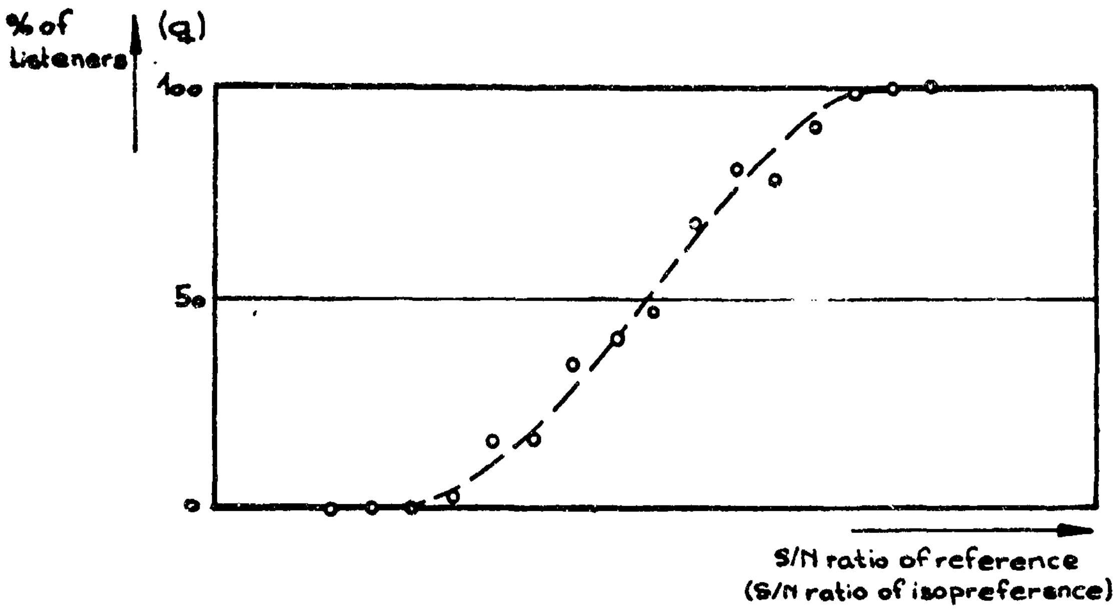


Fig. 2.22

distribution the parameters of which are called μ and σ. It is reasonable to take as well the mean value μ obtained in this manner as an approximate S/N ratio of isopreference for the group of listeners under test. The mean value μ will not differ very much from the mean value μ_q discussed above. But the standard deviation σ obtained now will be quite different to σ_q defined above. It will be the aim of the following paragraph to give the relation between these two standard deviations.

In the preceding section we have spoken about a distribution for the S/N ratios of isopreference evaluable for individual listeners provided a certain number of trials has been performed. In this connexion we shall accept normal distributions for these S/N ratios of isopreference with equal standard deviations σ_i for each individual listener. The mean values of these distributions may differ of course corresponding to the distribution of the S/N ratios of isopreference concerning the population of listeners assumed previously. Based on these premises it can be stated that the standard deviation σ

obtained by evaluating the S/N ratio of isopreference directly from the votes of listeners under test will be larger than the standard deviation σ_g obtained by evaluating the S/N ratios of isopreference of the individual listeners. The following equation holds /6/ :

$$\sigma = \sqrt{\sigma_g^2 + \sigma_i^2} \tag{2.6}$$

The increase of the standard deviation by evaluating the total votes of listeners at each S/N ratio is obvious, if one considers that certain "fail" votes of listeners can be eliminated by pre-evaluating S/N ratios of isopreference for single listeners.

In this section we have given two principle methods for evaluating the data collected in preference tests. Both methods lead to more or less the same S/N ratios of isopreference for groups of listeners, whereas the standard deviations for the approximating normal distributions are different. Since we are mainly interested in the S/N ratio of isopreference for groups of listeners we may take advantage of the method which requires less effort.

2.42 Processing of Test Data

In the course of preference tests reference signals B having several S/N ratios are presented to a group of listeners who decide at each S/N ratio whether they prefer the reference signal B to the test signal A. The votes of the listeners under test are the collected data.

At first the percentage of votes is calculated for B at each S/N ratio of the reference. This percentage will be taken with respect to the total number of listeners in the group when the S/N ratio of isopreference is to be evaluated for the group. But the percentage will be taken only with respect to the number of presentations when

S/N ratios of isopreference are to be evaluated for individual listeners.

We assume these percentages just mentioned above to be approximate values of the distribution of the probability that a S/N ratio of isopreference is below the S/N ratio just considered. We take for granted that all considered sets of S/N ratios of isopreference shall be normally distributed. The percentages of votes for signal B which we obtain in preference tests in dependence on several S/N ratios of reference signals will come close to the assumed distribution function. We may approximate the latter function by plotting a smooth curve between the points given by the percentages of votes versus the S/N ratios of the reference signals. (Fig. 2.23). This smooth curve shall be calculated as a normal distribution function with the mean value μ and the standard deviation σ

$$\Phi(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \int_{-\infty}^{x} \exp\left[-\frac{(\xi-\mu)^2}{2\sigma^2}\right] d\xi \qquad (2.7)$$

Lateron, wherever necessary we shall distinguish by subscripts to the parameters μ and σ between approximations to different distributions. There are the distributions of the S/N ratios of the individual listeners (μ_i, σ_i), those of groups calculated from the single S/N ratio of individual listeners (μ_g, σ_g), and those of groups calculated directly from the votes of individual listeners (μ, σ), (Sec. 2.41) The evaluation of a proper $\Phi(x,\mu,\sigma)$ is performed by the concept of the least mean square error defined by

$$F(\mu,\sigma) = \sum_{j=1}^{m} \left[y(x_j) - \Phi(x_j,\mu,\sigma)\right]^2 = \mathrm{Min.} \qquad (2.8)$$

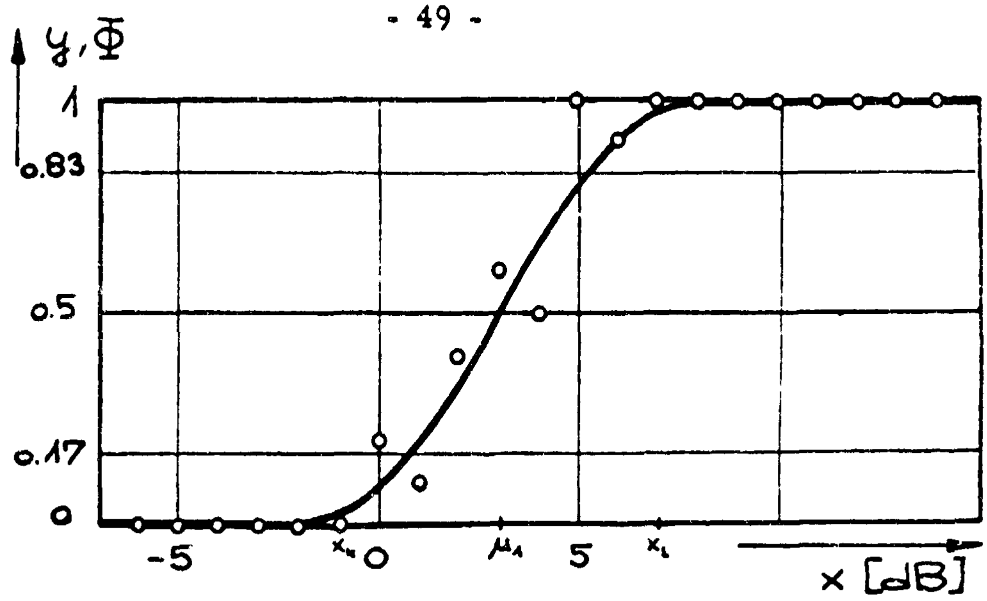


Fig. 2.23

In this equation $y(x_j)$ stands for the percentage of listeners voting for signal B presented with the S/N ratio of x_j dB. The sum is taken over all m presentations of a certain preference test.

The parameters μ and σ of approximating normal distributions will be evaluated by minimizing $F(\mu,\sigma)$. For that purpose an iteration procedure programmed on a digital computer may be used. The procedure starts with an approximate value for the desired mean value μ. The approximate value μ_1 should meet best the following equation

$$\sum_{j=1}^{\mu_1} y^2(x_j) - \sum_{j=\mu_1+1}^{m} [1-y(x_j)]^2 = \emptyset \qquad (2.9)$$

Additionally an approximate value σ_1 for the desired standard deviation σ has to be assumed. x_k shall denote the highest S/N ratio of the reference where all votes are for signal A, and x_ℓ shall denote the lowest S/N ratio of the reference where all votes are for signal B. Then we assume

$$\sigma_1 = \frac{x_l - x_k}{4} \qquad (2.10)$$

Experience proved the practicality of Equ. 2.9 and Equ. 2.10. They give for the example shown in Fig. 2.23 μ_1 = 3 dB and σ_1 = 2 dB. Now we take the normal distribution $\Phi(x, \mu_1, \sigma_1)$ as a first approximation and evaluate the mean square error $F(\mu_1, \sigma_1)$ defined by Equ. 2.8 for the given test data. Now for varying the parameters of $\Phi(x, \mu_1, \sigma_1)$ we take the following eight couples of parameters :

$$(\mu_1 \pm \Delta_1, \sigma_1) \;,\; (\mu . \pm \Delta_1, \sigma_1 . e^{\pm \delta_1}), (\mu_1, \sigma_1 . e^{\pm \delta_1})$$

with Δ_1= 1 dB and δ_1 = 1. After calculating the eight mean square errors for these distributions they are compared with $F(\mu_1, \sigma_1)$. If one of them is smaller than $F(\mu_1, \sigma_1)$ it is called $F(\mu_2, \sigma_2)$ and μ_2 and σ_2 are used for the next iteration step. But if $F(\mu_1, \sigma_1)$ is still the smallest mean square error we shall vary the couple of parameters (μ_1, σ_1) as follows :

$$(\mu_1 \pm \Delta_2, \sigma_1), (\mu_1 \pm \Delta_2, \sigma_1 . e^{\pm \delta_2}), (\mu_1, \sigma_1 . e^{\pm \delta_2})$$

with $\Delta_2 = \frac{\Delta_1}{2}$ and $\delta_2 = \frac{\delta_1}{2}$ and repeat the above procedure. This process converges rather rapidly to the desired parameters μ and σ. The iteration procedure is stopped when Δ drops below $2^{-4}\Delta_1$. This corresponds to an accuracy for μ better than 0,1 dB. For the actual numerical processing the normal distribution function $\Phi(x, \mu, \sigma)$ is calculated by means of an approximation formula $\overline{\Phi}(x, \mu, \sigma)$ which is sufficiently accurate for any x /8/ with

$$\overline{\Phi}(x, \mu, \sigma) = \begin{cases} \dfrac{1}{2(1 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4)^4} & z \leqq 0 \\[2ex] 1 - \dfrac{1}{2(1 + a_1 z + a_2 z^2 + a_3 z^3 + a_4 z^4)^4} & z \geqq 0 \end{cases}$$

and $$z = \frac{x - \mu}{\sigma}$$

$$a_1 = 0,278393$$
$$a_2 = 0,230389$$
$$a_3 = 0,000972$$
$$a_4 = 0,078108$$

The procedure described above enables us to fit a cumulative normal distribution to data which consist in given frequencies y_i over S/N ratios x_i. Another task was described in the preceding chapter, the calculation of the S/N ratio of isopreference μ_g for a group of listeners from the μ_i for each listener out of this group. These μ_i can also be calculated by the method given above. The mean value μ_g can be calculated simply by

$$\mu_g = \frac{1}{N} \cdot \sum_{i=1}^{N} \mu_i$$

where N stands for the number of listeners in the group. The corresponding standard deviation is found by

$$\sigma_g = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\mu_i - \mu_g)^2}$$

In all cases discussed so far statistical data were evaluated, i.e. many or at least some similar data were available. Beyond this the question may arise whether one can find the point of isopreference for a single listener from his decisions given in a single test run. Of course this is a rather poor basis for a "statistical" evaluation. An estimate which proved to be useful may be found in the following way. It shall be supposed that the set of presented reference signals has equidistant S/N ratios. Some examples of possible decision series are shown in Fig. 2.24.

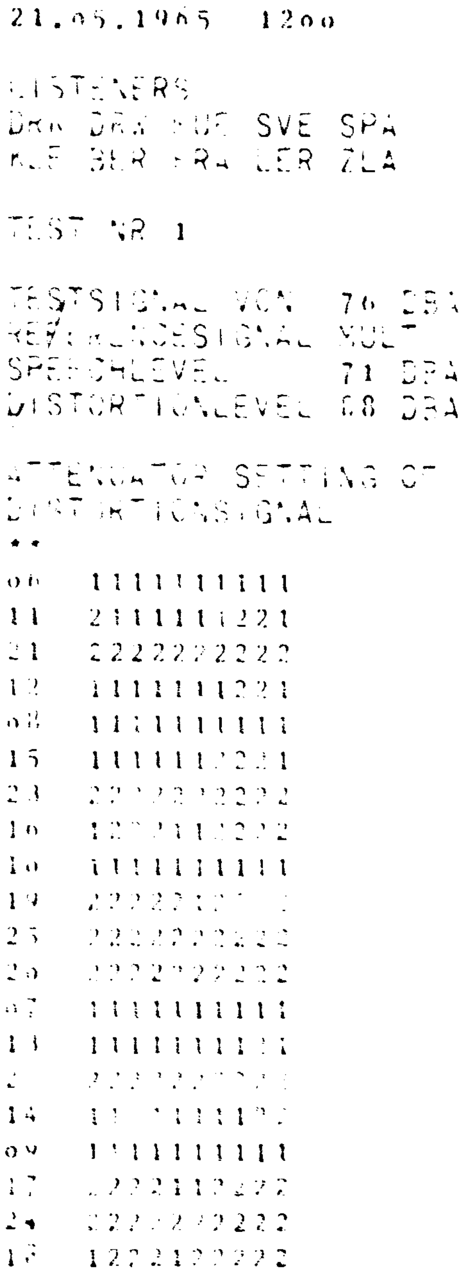
S/N ratio[dB]	1	2	3	4	5	6	7	8	9	10	11	12	13	14
I	A	A	A	A	A	A	A	B	B	B	B	B	B	B
II	A	A	A	A	A	A	B	A	B	B	B	B	B	B
III	A	A	B	A	B	A	A	A	A	A	B	B	B	B

Fig. 2.24

The point of isopreference is to be expected to lie between the limits of consistent decisions. The mean value is easily found in series I to be 7.5 dB. This holds as well for series II because of the symmetrical decisions. It is rather difficult to define an isopreference level for series III. The proposed solution is to have as many "displaced" A decisions as "displaced" B decisions on either side of the thus determined point of isopreference M.

The described procedures have been programmed in FORTRAN for processing on an IBM 7040 digital computer system. The programs and sample printouts are contained in the Appendix.

The input data for these programs are obtained and stored in the following formats : the decisions given by the listeners are stored by means of the teletypewriter as shown in Fig. 2.25. The numbers printed at the left side give the respective test condition by the attenuator setting for the distortion signal. The according listeners' decisions are printed automatically on the right. The numerals 1 or 2 stand for a decision preferring the corresponding signal A or B. Test conditions, e. g. signal specifications, date, and listener names, have to be written by the operator. The decisions are ranked and prepared for further handling by the test operator,while the next test is running. This turned out to be very useful, because one can control the listeners and the presented test conditions throughout the test. The form used for this purpose is shown in Fig. 2.26. The markers in this table give the reference conditions where the respective listener preferred signal B and are very easy to survey. Cards are punched from these data in a format which is specified in the Appendix and fits the requirements of our FORTRAN programs.



```
21.05.1965   1200

LISTENERS
DRK DRU MUE SVE SPA
KLE BER FRA LER ZLA

TEST NR 1

TESTSIGNAL VON   76 DBA
REFERENCESIGNAL MULT
SPEECHLEVEL      71 DBA
DISTORTIONLEVEL  68 DBA

ATTENUATOR SETTING OF
DISTORTIONSIGNAL
**
06   1111111111
11   2111111221
21   2222222222
12   1111111221
08   1111111111
15   1111112221
23   2222222222
16   1222112222
10   1111111111
19   2222212222
25   2222222222
20   2222222222
07   1111111111
13   1111111121
22   2222222222
14   1121111122
09   1111111111
17   2222112222
24   2222222222
18   1222122222
```

Fig. 2.25

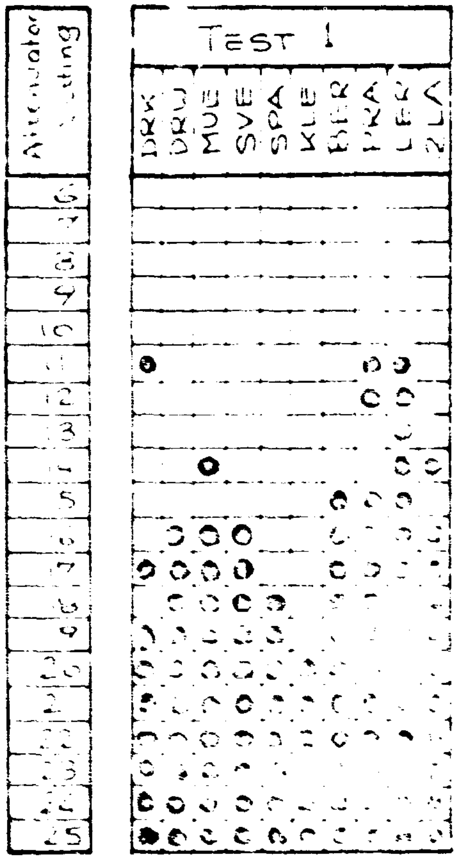

Attenuator setting
Test 1
DRK
DRU
MUE
SVE
SPA
KLE
BER
FRA
LER
ZLA
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25

Fig. 2.26

this purpose is shown in Fig. 2.26. The markers in this table give the reference conditions where the respective listener preferred signal B and are very easy to survey. Cards are punched from these data in a format which is specified in the Appendix and fits the requirements of our FORTRAN programs.

```
21.o5.1965   12oo

LISTENERS
DRK DRW MUE SVE SPA
KLE BER PRA LER ZLA

TEST NR 1

TESTSIGNAL VON    76 DBA
REFERENCESIGNAL  MULT
SPEECHLEVEL       71 DBA
DISTORTIONLEVEL  88 DBA

ATTENUATOR SETTING OF
DISTORTIONSIGNAL
**
o6    1111111111
11    2111111221
21    2222222222
12    1111111221
o8    1111111111
15    1111112221
23    2222222222
16    1222112222
1o    1111111111
19    2222212222
25    2222222222
2o    2222222222
o7    1111111111
13    1111111121
22    2222222222
14    1121:11122
o9    1111111111
17    2222112222
24    2222222222
18    1222122222
```

Attenuator Setting	Test 1									
	DRK	DRW	MUE	SVE	SPA	KLE	BER	PRA	LER	ZLA
6										
7										
8										
9										
10										
11	○							○	○	
12								○	○	
13									○	
14			○						○	○
15							○	○	○	
16		○	○	○			○	○	○	○
17	○	○	○	○			○	○	○	○
18		○	○	○	○		○	○	○	○
19	○	○	○	○	○		○	○	○	○
2o	○	○	○	○	○	○	○	○	○	○
21	○	○	○	○	○	○	○	○	○	○
22	○	○	○	○	○	○	○	○	○	○
23	○	○	○	○	○	○	○	○	○	○
24	○	○	○	○	○	○	○	○	○	○
25	○	○	○	○	○	○	○	○	○	○

3. EXPERIMENTAL EVALUATION OF PREFERENCE TESTING

3.1 Scope of Measurements

In this section under the subtitles of speech quality, intelligibility, loudness and human factors a sequence of separate topics will be treated. We have not tried to fit all those topics into one large picture, because we felt that until now in some critical areas we do not have enough data to establish final statements.

3.11 Speech Quality

Repeated preference tests

For studying the behavior of listeners preference tests were performed under equivalent conditions with several groups of listeners. Tests were repeated during a session with different sequences of the reference. The normal vocoder (VON) was used as test signal and hifi speech signals degraded by multiplicative noise (MULT) or by additive noise (ADD) as reference signals.

The S/N ratio of the reference signal for isopreference will be called isopreference level in the following. The isopreference levels were determined for the individual listeners by analyzing their votes separately for all tests performed during one test session, as well as those for each group of listeners at each test run. Further points of isopreference for single listeners and for groups were obtained by evaluating their total votes during one session. By comparison of these isopreference levels calculated from the group data and from those of single listeners it could be confirmed that the relation holds between the standard deviations as given by Equ. 2.6.

The results of tests performed with one group of listeners on one day (21-5-65) shall be given now. Further results for two other groups of listeners are given in the Appendix.

A number of 10 listeners compared the normal vocoder (VON) to the multiplicative reference. The test was repeated 6 times within a two hour session.

At first examples shall be given of the votes of the 10 listeners in one test and of the votes of a single listener in the 6 consecutive test runs of the session. In the following tables the S/N ratios of the reference signals presented are listed and the votes of listeners favoring the reference signal are marked. The listeners are named by capital letters. The complete data for the session are listed in the Appendix.

Attenuator Setting	TEST 1									
	DRK	DRU	MUE	SVE	SPA	KLE	BER	PRA	LER	ZLA
6										
7										
8										
9										
10										
11	●							●	●	
12								●	●	
13									●	
14			●						●	●
15							●	●	●	
16		●	●	●			●	●	●	●
17	●	●	●	●			●	●	●	●
18		●	●	●	●		●	●	●	●
19	●	●	●	●	●		●	●	●	●
20	●	●	●	●	●	●	●	●	●	●
21	●	●	●	●	●	●	●	●	●	●
22	●	●	●	●	●	●	●	●	●	●
23	●	●	●	●	●	●	●	●	●	●
24	●	●	●	●	●	●	●	●	●	●
25	●	●	●	●	●	●	●	●	●	●

Attenuator Setting	DRK					
	Test 1	Test 2	Test 3	Test 4	Test 5	Test 6
6						
7						
8						
9						
10						
11	●					
12						
13						
14						
15						
16		●		●		
17	●	●	●		●	
18		●	●	●	●	
19	●	●	●			
20	●	●	●	●	●	●
21	●	●	●	●	●	●
22	●	●	●	●	●	●
23	●	●	●	●	●	●
24	●	●	●	●	●	●
25	●	●	●	●	●	●

Fig. 3.1

From these data we obtained the isopreference levels μ_i for the single listeners and the standard deviations σ_i of the approximating normal distributions. Examples for such normal distributions are given in the Appendix.

Listener		DRK	DRU	MUE	SVE	SPA	KLE	BER	PRA	LER	ZL
μ_i	dB	0.2	-0.5	-2	-1.6	0.3	1.6	-1.6	-1.3	-3.0	-2.2
σ_i	dB	2.2	1.6	1.8	1.7	1.1	1.5	2.1	1.6	2.3	1.2

Table 3.1

A schematic diagram of these results is shown in Fig. 3.2. The mean square of the standard deviations σ_i for the 10 listeners is found to be

$$\overline{\sigma_i} = 1,75 \text{ dB}$$

The isopreference levels of the group for each test run are listed in Table 3.2 and schematically shown in Fig. 3.3. Examples for normal distributions approximating the group decisions in separate test runs are given in the Appendix.

TEST NR		1	2	3	4	5	6
μ	dB	-1.7	-1.5	-2.7	0	-0.4	-0.5
σ	dB	2.8	2.0	2.2	2.3	0.8	2.7

Table 3.2

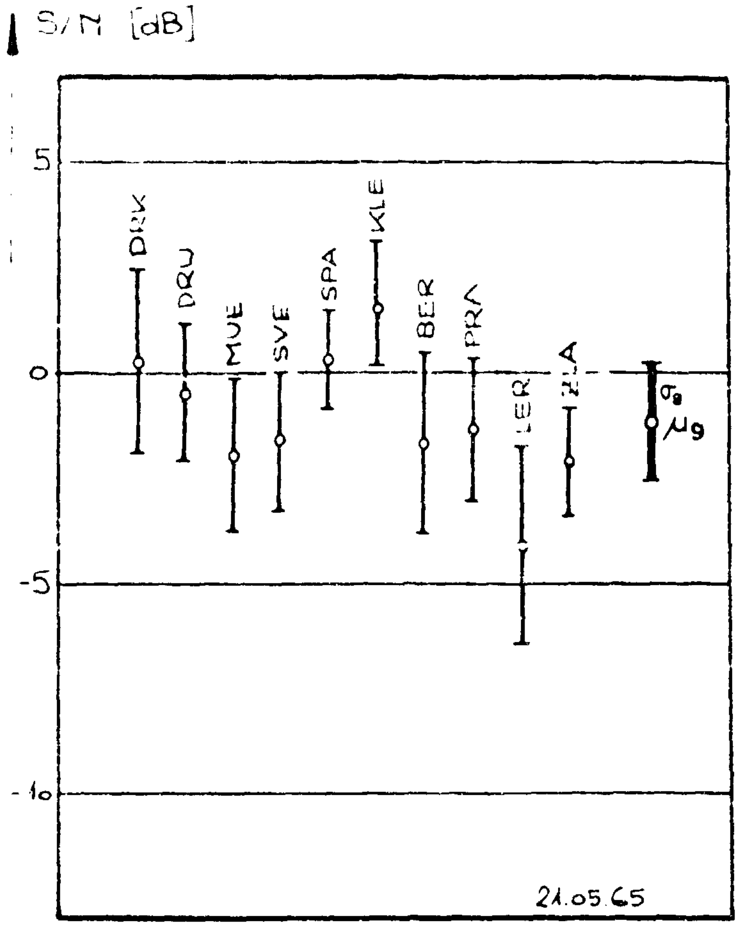


Fig. 3.2

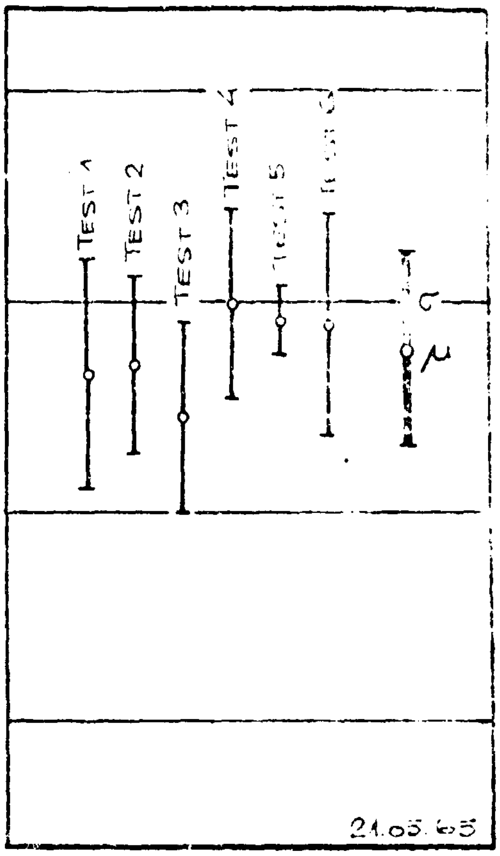


Fig. 3.3

The isopreference level for the group over the whole session was not only calculated as μ_g from the isoprefe ence levels of the individual listeners which are listed as μ_i in Table 3.1 but also from the lump sum of all decisions in the session denoted as μ .

$$\mu_g = -1.0 \text{ dB} \qquad \mu = -1.1 \text{ dB}$$
$$\sigma_g = 1.3 \text{ dB} \qquad \sigma = 2.3 \text{ dB}$$

These results show good conformity. From the standard deviations σ_g and σ one may deduce an estimate for the standard deviation σ_{iest} of the decisions of individual listeners in the session. We obtain an estimate by using Equ. 2.6

$$\sigma_{i\,est} = \sqrt{\sigma^2 - \sigma_g^2} = 1{,}9\,dB$$

This estimate value is found to be close to $\overline{\sigma_i}$ = 1,75 as calculated above. These results and those of all similar tests may be summarized : the iso-preference levels for individual listeners vary over a range of about 10 dB. The respective standard deviations were all around 2 dB. At the same time the uncertainty range of individual listeners was also found to be about ± 2 dB.

A typical preference test session

The following example shows the results from a complete test session. A number of 6 well trained listeners made judgements on four test systems HP, LP, VON, and VO2 in comparison with the additive and multiplicative reference. The Table 3.3 shows the mean value M of the single listener for single test runs and in the last two columns the mean value μ and the standard deviation σ for the whole group are given.

Listener		PRA		ZLA		STE		SVE		MUE		DRW		GROUP	
Signal															
Test	Ref	M	direct judgem	M		M		M		M		M		μ	σ
HP	ADD	5.5	1	7.5	1	5.5	1	4.5	1	5.5	2	7.5	1	6.5	1.5
	MULT	2.5		4.5		1.5		0.5		2.5		3.5		2.5	1.8
LP	ADD	4.5	2	2.5	2	5.5	2	5.5	2	4.5	1	4.5	2	4.6	1.8
	MULT	0.5		1.5		0.5		-0.5		4.5		4.5		2.0	2.0
VO1	ADD	3.5	3	1.5	3	3.5	3	2.5	3	0.5	3	2.5	3	2.1	1.6
	MULT	-1.5		-1.5		-2.5		-3.5		-0.5		0.5		-1.4	1.6
VO2	ADD	0.5	4	0.5	4	1.5	4	0.5	4	-3.5	4	-0.5	4	0	2.0
	MULT	-5.5		-10.5		-5.5		-8.5		-9.5		-9.5		-8	2.4

Table 3.3

M and μ are given in terms of S/N ratios. Therefore the highest value represents the test signal with the best quality and the lowest value belongs to the worst test signal.

After these normal preference tests, the listeners were requested to rank order the four now well known signals by giving marks between one and four from the best to the worst quality signal. The results from these direct judgements of the listeners can be found in the respective columns beside the mean values M from the preference

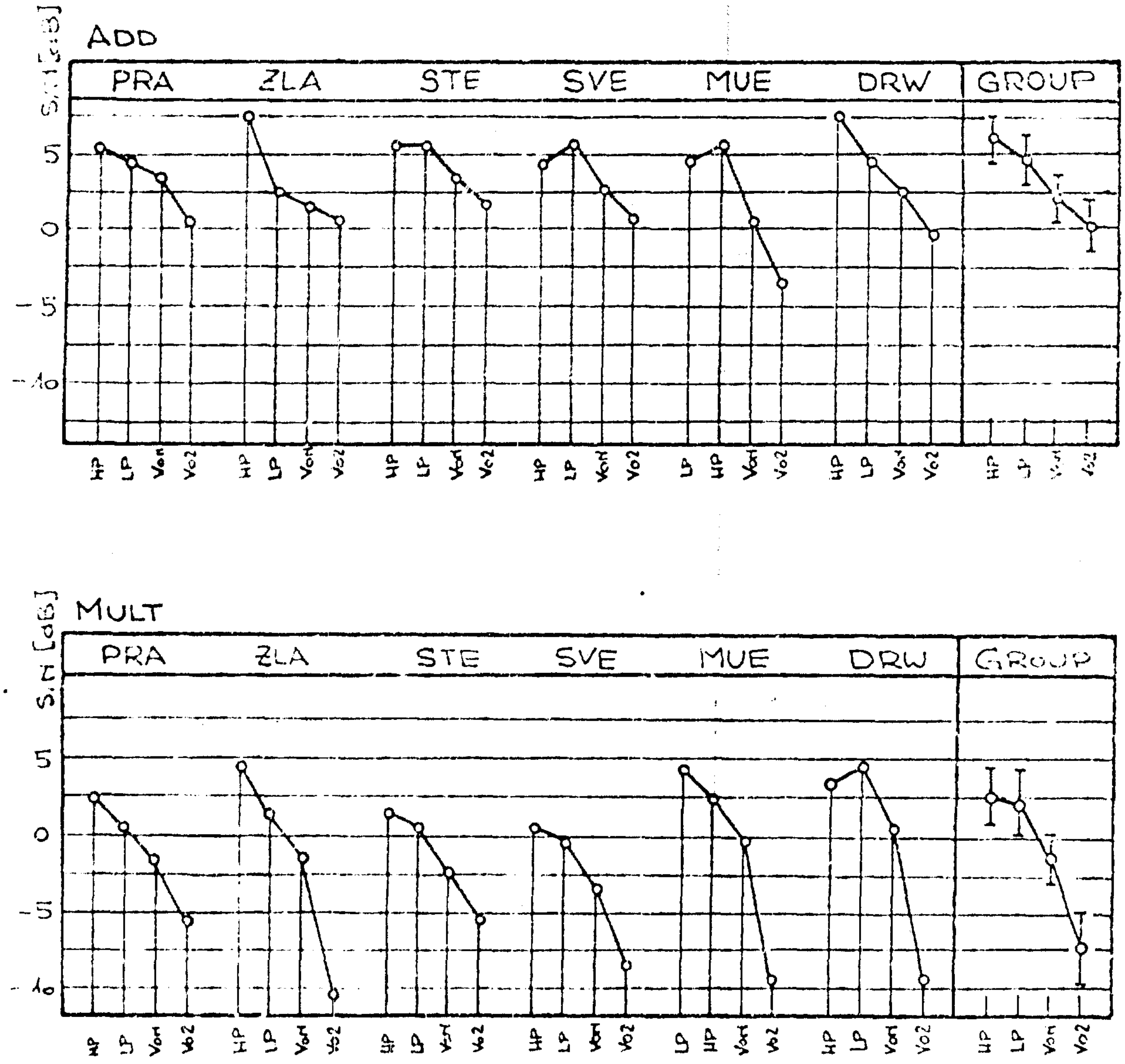


Fig. 3.4

tests. With the exception of the listener MUE all others have the order HP - LP - VON - VO2 from the best to the worst signal when asked directly. This should coincide with the preference results expressed in values of M.

The values M for both references and the mean values μ of the group are plotted in Fig. 3.4. The monotonous decrease of M for the first three listeners shows that the rank order is the same for both direct and preference comparisons. The results from the last three listeners are in two cases with the additive reference and in one case with the multiplicative reference contradicting the results of the direct judgements. The mean values μ of the group are in all cases in the same order. The standard deviations σ are relatively small with an average value of about 1,8 dB. The same test signals HP, LP, VON and VO2 were also judged by the group of 38 untrained listeners. For each listener the isopreference levels regarding each of the eight preference judgements performed were evaluated and their distribution was plotted. Fig. 3.5 shows such a distribution for a comparison of VON and MULT.

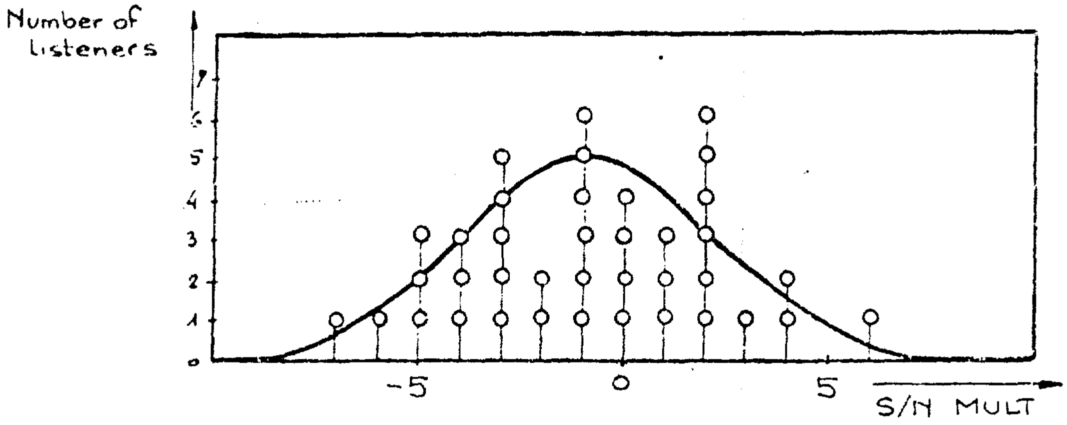


Fig. 3.5

These distributions can be characterized by mean values μ and standard deviations σ derived from their first and second moments. For the eight test runs in this connexion the values μ_g and σ_g are given in Table 3.4 and in Fig. 3.6.

		MULT				ADD			
		Vo2	VoN	LP	HP	Vo2	VoN	LP	HP
μ_g	dB	-3.6	-1	3.5	8.1	3	4.4	6.5	9.3
σ_g	dB	4	3.2	5.9	4.2	3	3.5	7	4.4

Table 3.4

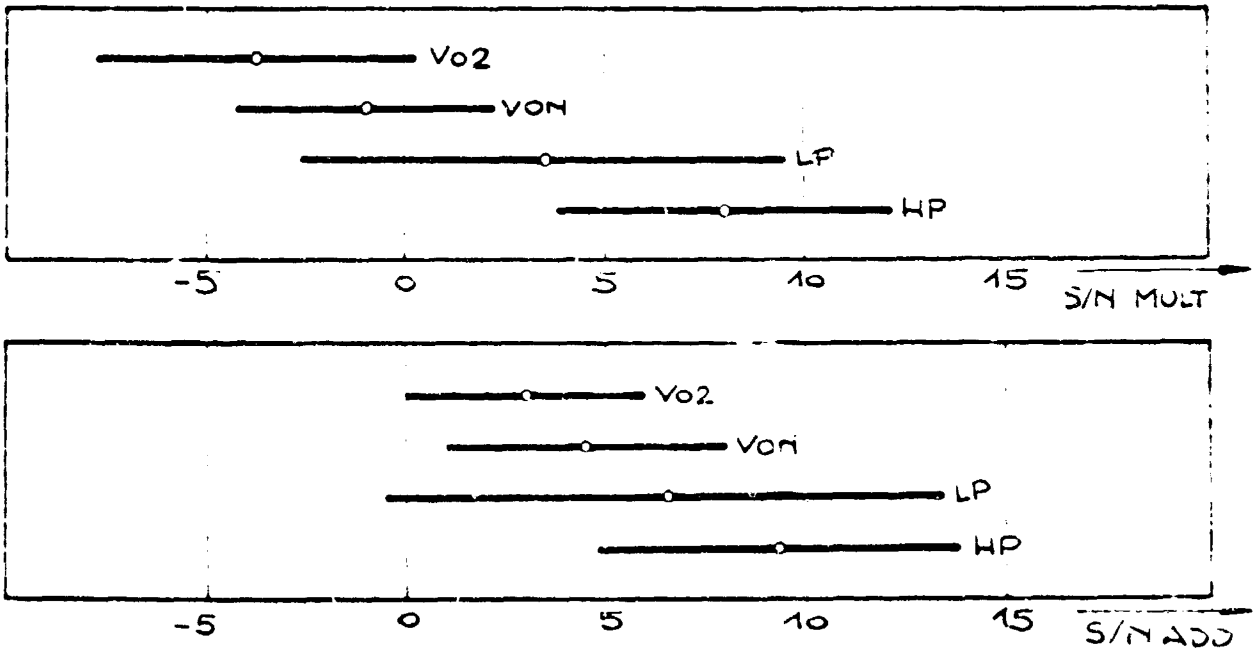


Fig. 3.6

The normal density functions with the mean values μ_g and standard deviations σ_g for the eight test cases are shown in Fig. 3.7 on the next page.

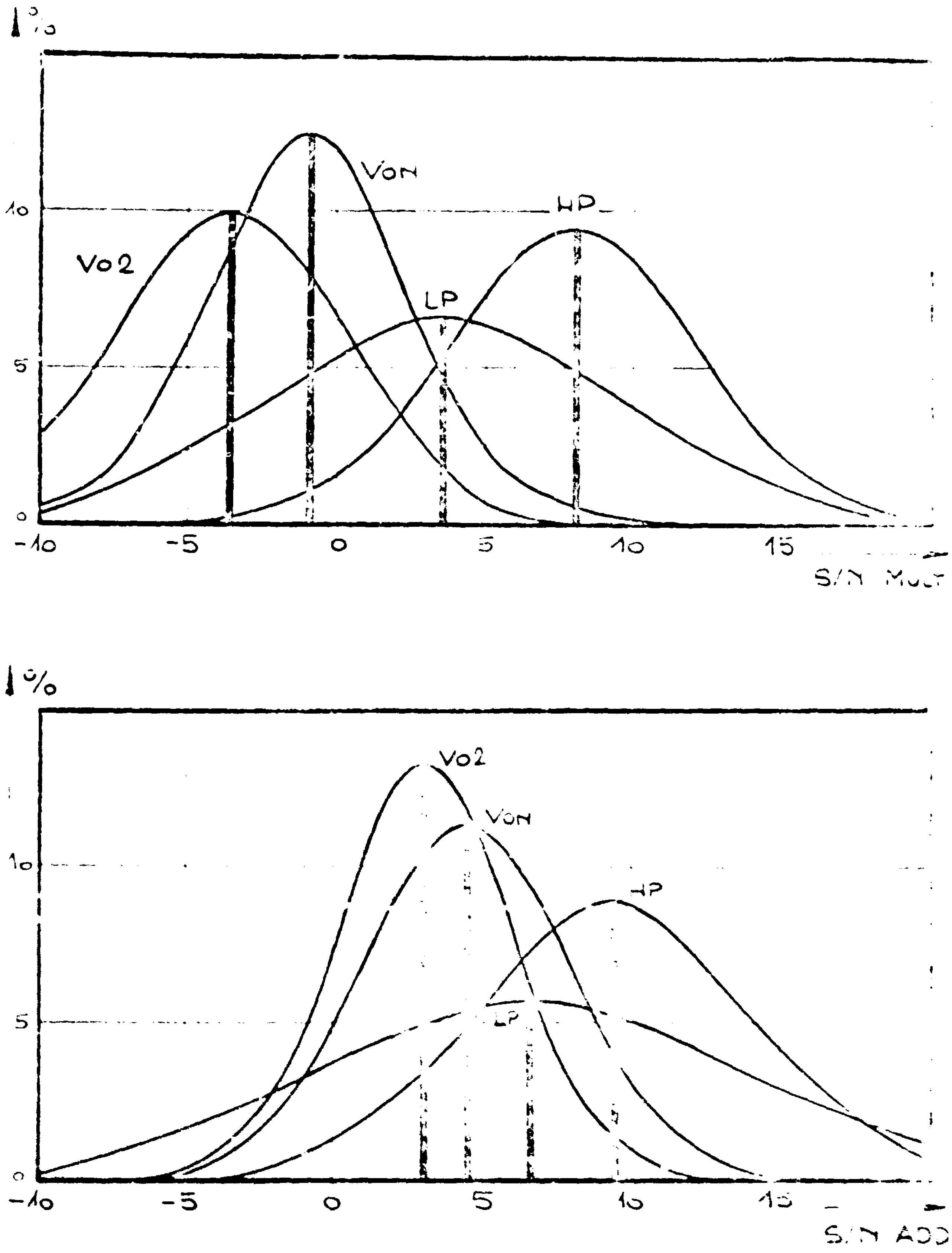

%
10
5
0
-10
-5
0
5
10
15
S/N MULT
Vo2
VoN
LP
HP
%
10
5
0
-10
-5
0
5
10
15
S/N ADD
Vo2
VoN
HP
LP

Fig. 3.7

Correspondence between the additive and the multiplicative reference

As we have been working mainly with the two references ADD and MULT, we tried to find a correspondence between them. The relation has been studied for its reproducibility not only when executing the same test on different days and with different listeners, but also when different test procedures are utilized for its determination.

At first pair comparison tests were made with both signals varied simultaneously in quality. All these speech pairs consisting of an ADD signal with any S/N ratio compared with a MULT signal with any other S/N ratio have been presented to the listeners in totally random order. This test procedure is different from our normal preference test procedure.

The [illegible] this new procedure can be plotted in a three dimensional system with the two horizontal axis numbered in S/N ratios or attenuator settings of ADD and MULT, and along the third vertical axis percentages of the listeners are given who prefer ADD. This plotting procedure would yield the relation between the two signals as a three dimensional surface. With regards to the difficulties of using a three dimensional plotting scheme, it was decided to use only two dimensional mapping. The presented signal pairs correspond then to points of the area between the two reference axis and are labelled with the respective percentage of listeners preferring ADD. We have decided in view of this mapping to call the new test procedure "area test".

The curve of intersection between the three dimensional surface and a horizontal plane at 50 % listeners preference represents the "isopreference curve" of the two signals ADD and MULT under test.

In a real test there will be some'wrong' decisions in the critical range close to the isopreference curve. In order to determine the isopreference curve cross cuts perpendicular to the ADD and MULT axis of the data surface are made. In these cross cuts the isopreference level discussed together with a standard deviation is calculated as in Sec. 2.4. The desired isopreference curve will then be found as an empirical approximation to those individually determined isopreference points.

As an example the results of such an area test shall be given which has been conducted of 10 listeners. Fig. 3.8 shows the decisions for the listener SCW and Fig. 3.9 shows the results for the whole group. Fig. 3.9 is a computer printout where the preferences of the listeners are mapped in the respective test points as explained before. The results of the evaluation of both groups of distribution functions, or cross cut approximations, as explained above, are plotted ing Fig. 3.10. The mean values of both groups of functions are discriminated by circles and triangles. The isopreference curve approximates these points. Fig. 3.10 gives also the corresponding standard deviations σ of the distribution functions.

The reproducibility of these measurements is demonstrated in Fig. 3.11. Five such area tests at different times have been carried out. The results are shown as curves 1 to 5. The deviations are so small that we feel entitled to call the mean of these curves the "standard" isopreference curve between our signals ADD and MULT. All measured isopreference curves were found to deviate less than about ± 2 dB from the standard curve over a range of 20 dB of the multiplication reference. This expected uncertainty range of ± 2 dB is indicated by dashed lines in Fig. 3.11. When the quality of the test signals is increasing towards hifi quality, the deviations grow larger.

Parallel to the area tests also normal preference tests have been conducted in order to check and to verify the validity of the standard

```
MEASUREMENTS OF SPEECH QUALITY

SIGNAL A = 71 DBA HIFI / (88-ATTSETT) DBA MULT
SIGNAL B = 71 DBA HIFI / (90-ATTSETT) DBA ADD

AREA TEST
PERFORMED  2. 7. 1965
10 LISTENERS
DECISIONS OF LISTENER -SCH-

ATTENUATOR SETTING B
***
40.                         2 2 2 2 1 1

                      2   2 2 2   1 1 2 1

                        2 2 1 2 2 1 1 1 1

                      2 2   2 2 2 2 1 1

                ای 
```

```
MEASUREMENTS OF SPEECH QUALITY

SIGNAL A = 71 DBA HIFI / (88-ATTSETT) DBA MULT
SIGNAL B = 71 DBA HIFI / (90-ATTSETT) DBA ADD

AREA TEST
PERFORMED  2. 7. 1965
10 LISTENERS
DECISIONS OF TOTAL GROUP

ATTENUATOR SETTING B
***
40.                         8 9 8 5 4 1

                      A   A A 9   6 1 2 0

                        A 9 8 9 5 2 3 1 0

                      9 A   A 5 8 2 0 0

                A A A A 9 A 6 5 2 2 0

30.           A A A 8 8 A 9 4 3 2 3 2 3 0

            A   A 8 9 8 5 2 4 0 4 1

            A A 9 A 7 8 9 3 4 2 0 1

            A A 9 9 7 7 6 1 0 0 0 0

          A A 9 5 7 6 3 3 1 0 0 0

20. A A 9 A 9 8 6 5 0 0 1 0 0

    9 9 8 9 3 2 2 0 0

    5 7 4 2 0 0 1

    2 0 0 0 0 0

    0 0 0

10.










 0.

   0.       10.       20.       30.       40.
                        ATTENUATOR SETTING A
```

Fig. 3.9

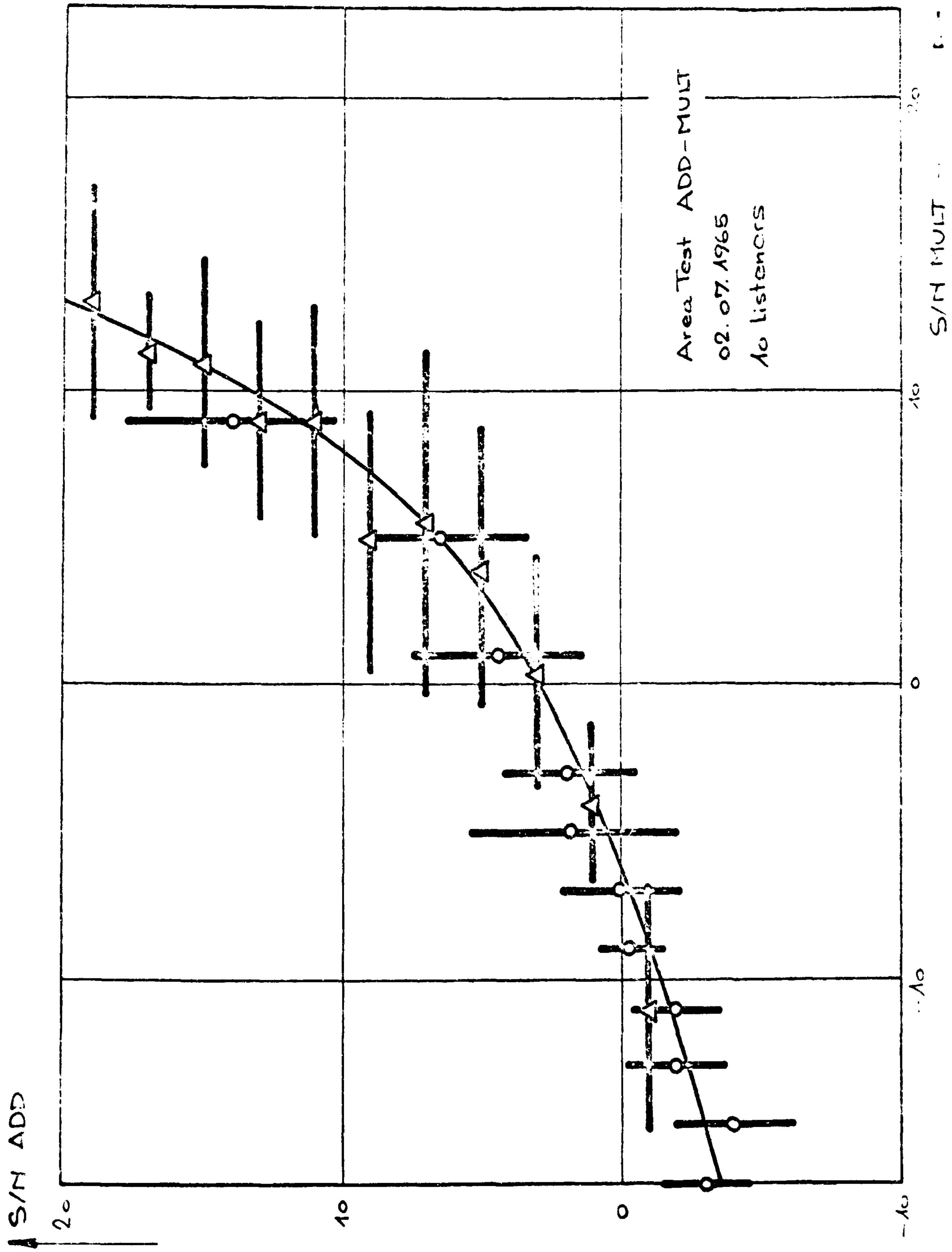

S/N ADD
20
10
0
-10
S/N MULT
20
10
0
-10
Area Test ADD-MULT
02.07.1965
10 listeners

Fig. 3.10

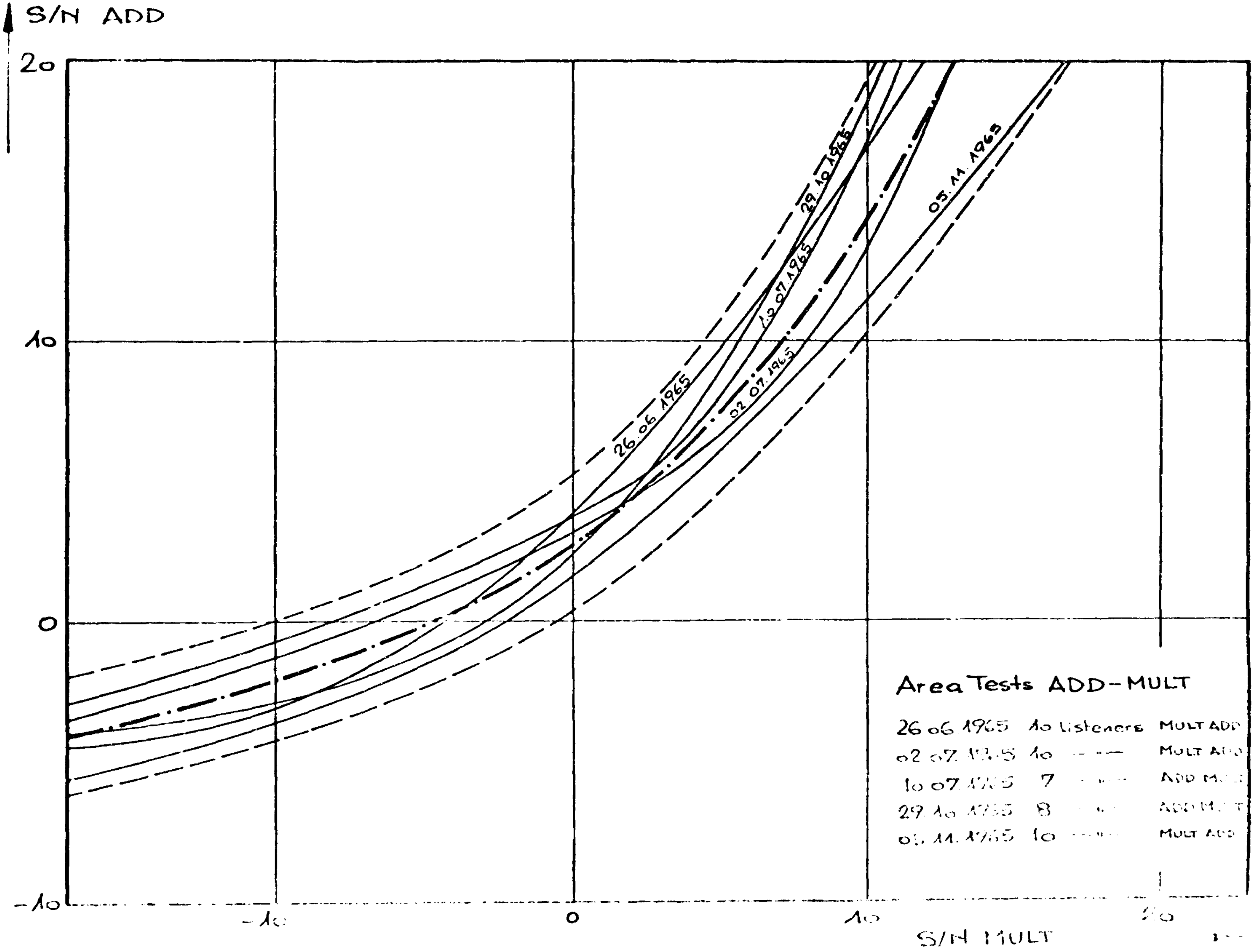

S/N ADD
20
10
0
-10
-10
0
10
20
S/N MULT
26.06.1965
29.10.1965
10.07.1965
02.07.1965
05.11.1965
Area Tests ADD-MULT
26.06.1965 10 listeners MULT ADD
02.07.1965 10 MULT ADD
10.07.1965 7 ADD MULT
29.10.1965 8 ADD MULT
05.11.1965 10 MULT ADD

Fig. 3.11

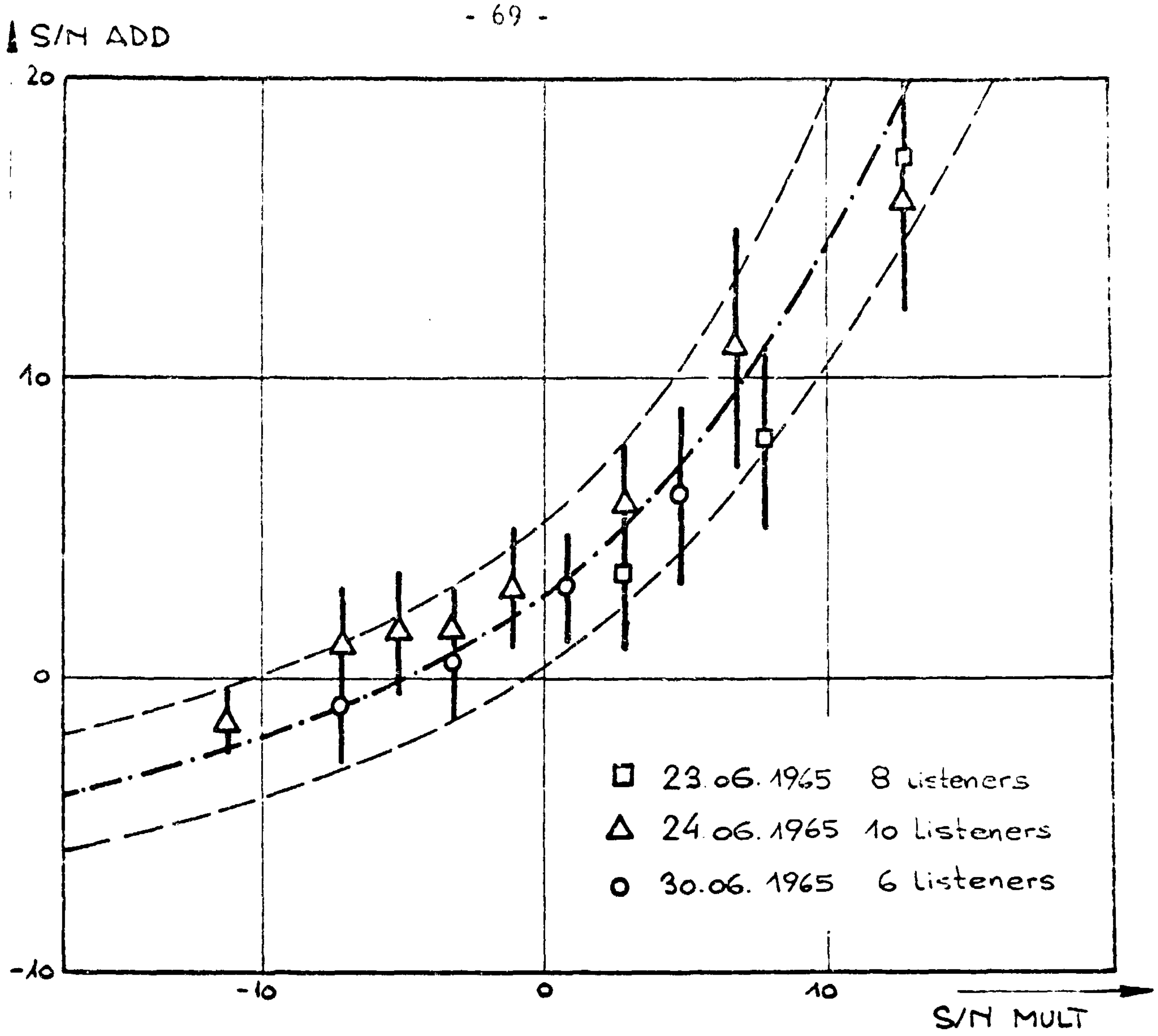


Fig. 3.12

isopreference curve. The two kinds of preference tests, possible in this connection, have been performed and repeated : the additive reference signal, held constant, acting therefore as the test signal and the multiplicative reference being varied, acting therefore as reference signal and vice versa. The corresponding figures Fig. 3.12 and Fig. 3.13 show the results for these two kinds of preference tests. Mean values and standard deviations for groups consisting of about 8 listeners are given. Both kinds of tests show in the middle range of the curve approximately the same σ values which are increasing for good or bad speech qualities. All isopreference points are lying within the uncertainty range defined above.

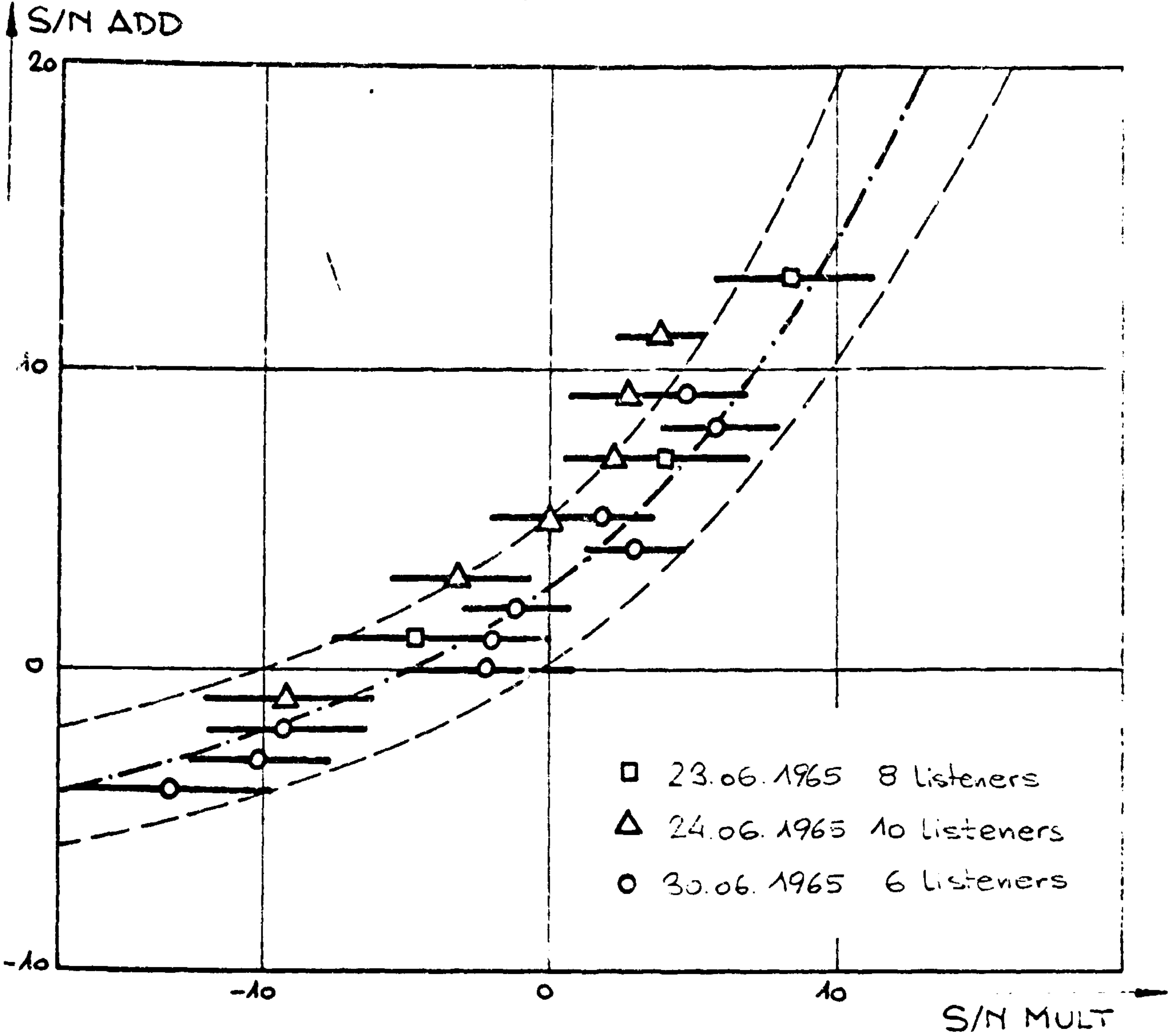


Fig. 3.13

Area tests with the digital reference

A few weeks ago after finishing an experimental set-up for the generation of the new "digital" reference DIG (Sec. 2.21) we started to work with the new signal. At first it was tried to establish the iso-preference relation to the other reference signals ADD and MULT. The results presented here are still preliminary, but fit quite well into the scope of our data.

Several area tests were made in the same way as described in the previous subsection. The results of the two tests with the signal pairs

DIG-ADD and DIG-MULT are shown in Fig. 3.14 and Fig. 3.15. The tests have been conducted twice at different days with 5 to 7 listeners.

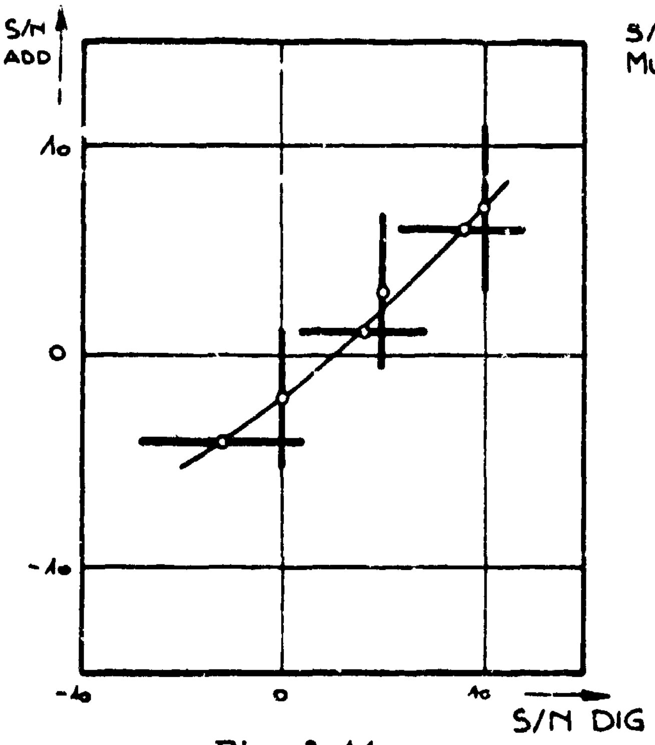


Fig. 3.14

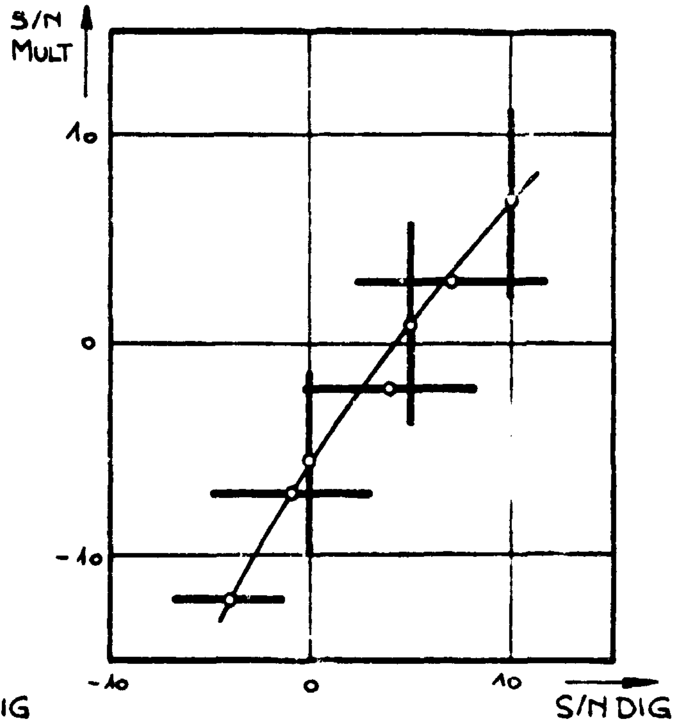


Fig. 3.15

The given standard deviations in both directions for the isopreference points seem to be somewhat larger than in the case of the ADD-MULT relation of Fig. 3.10, but should be confirmed by further tests.

A combination of the three relations ADD-DIG, DIG-MULT, MULT-ADD will allow to get information about the transitivity of preference measurements. This will be discussed in Sec. 3.2.

Area tests with distorted test signals

Normally one is not expected to be interested in degrading the test signals because one would want to test and to evaluate signals as they are in practical use. But we used this easy possibility to extend our speech material in order to get new dimensions and to prove already existing relations.

Several new test signals have been generated from the set of our test tapes and have been used for area tests. The results of four tests with the new test signals normal vocoder multiplied by noise VON-MULT, normal vocoder plus additive noise VON-ADD and highpass filtered speech plus additive noise HP-ADD in comparison with some of our reference signals are shown in Fig. 3.16 - 3.19. The test sessions have been attended by 7 to 10 listeners.

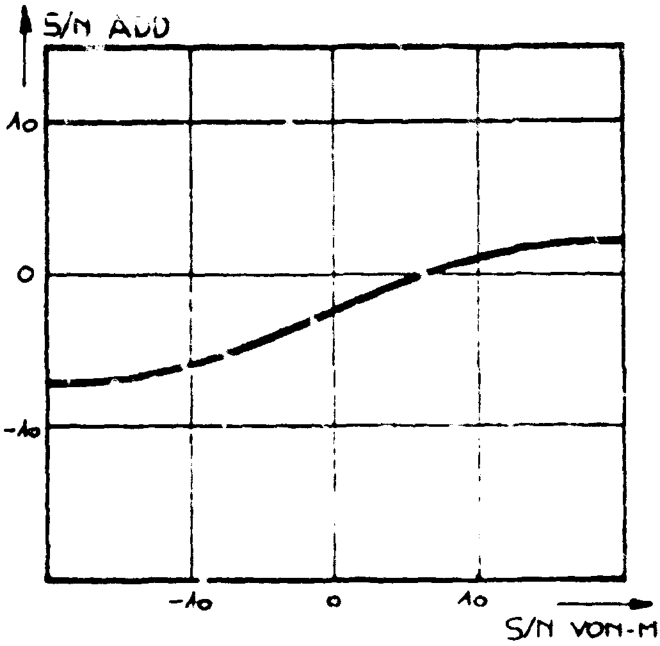


Fig. 3.16

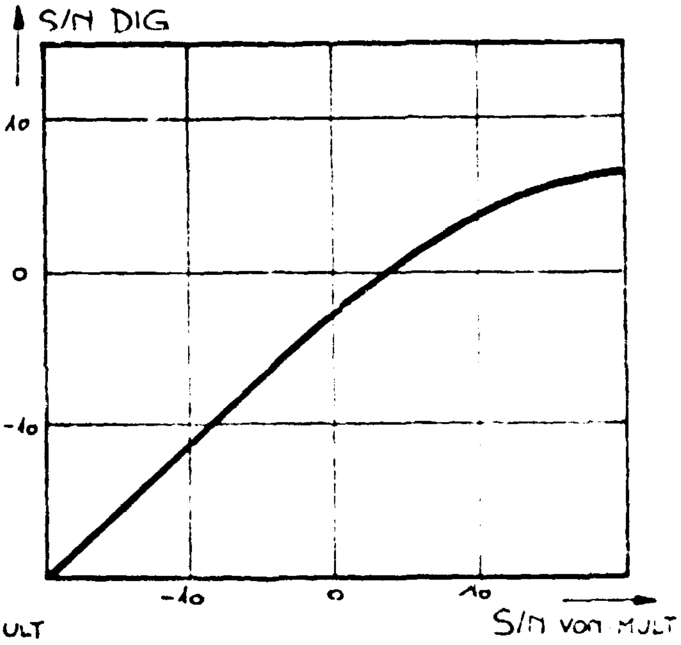


Fig. 3.17

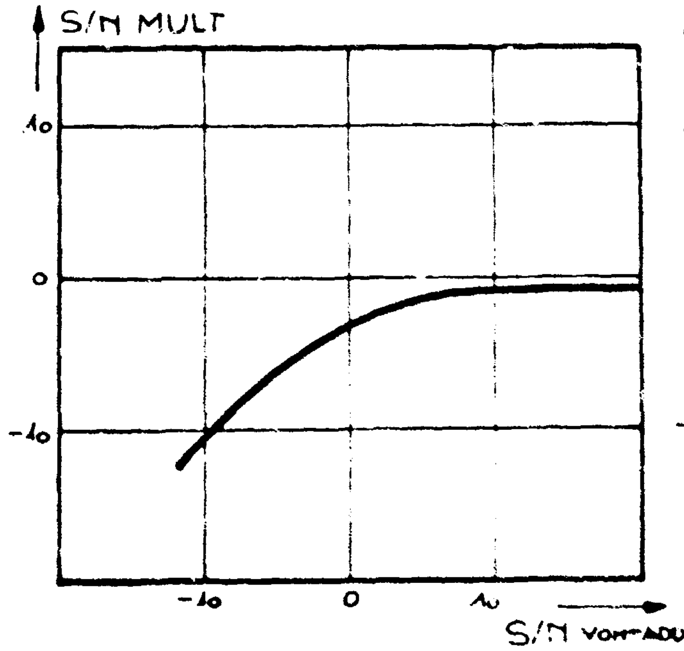


Fig. 3.18

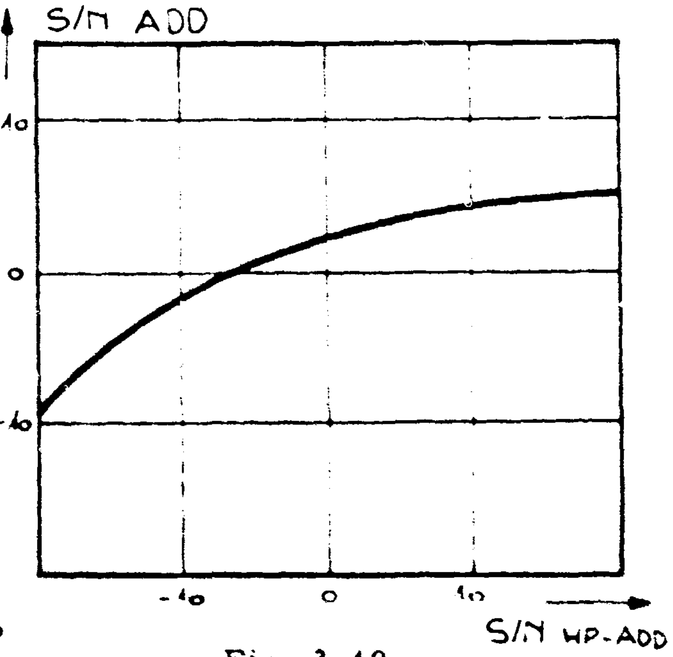


Fig. 3.19

Obviously, the diagrams have to show that for very high S/N ratios of the additional distoritons no change of the isopreference level of the test signal can be expected. But it is astonishing to see, e.g. in Fig. 3.18 that the quality of our vocoder speech signal could not be impaired by adding noise down to 6 dB S/N ratio.

This result shows that the influence of an additional degradation upon an already degraded speech signal may be difficult to predict. Effects of this kind will deserve further study. They may serve to give some guidance whether it will be worthwhile to improve single properties of a speech processing system under development. A parallel may be found in noise control work, where it is well known and easy to understand that isolated changes or reductions in a complex noise signal will have practically no effect on the overall loudness of this signal.

Rank order tests

Rank order tests have been started recently. Questions as the following shall be studied:

The ability of listeners to rank order systems with different S/N ratio,
the definition of difference limens in speech quality,
the relation between speech quality and S/N ratio of a speech signal.

Instead of signal pairs single speech signals with different S/N ratios are presented to the listeners. At first the listeners hear two reference signals, one with very good and the other with very bad quality. Then they are asked to rank order a random sequence of the above speech signals between the first two.

At first two tests have been carried out for DIG as a variable speech signal with groups of 7 listeners. The quality has been varied in 10 incremental steps of 2 dB S/N ratio covering a total range of 20 dB. The listeners were requested to classify each of nine signals after a single presentation of 5 seconds by giving marks between one and nine to these signals, while mark zero for the best and mark ten for the worst signal had been previously defined.

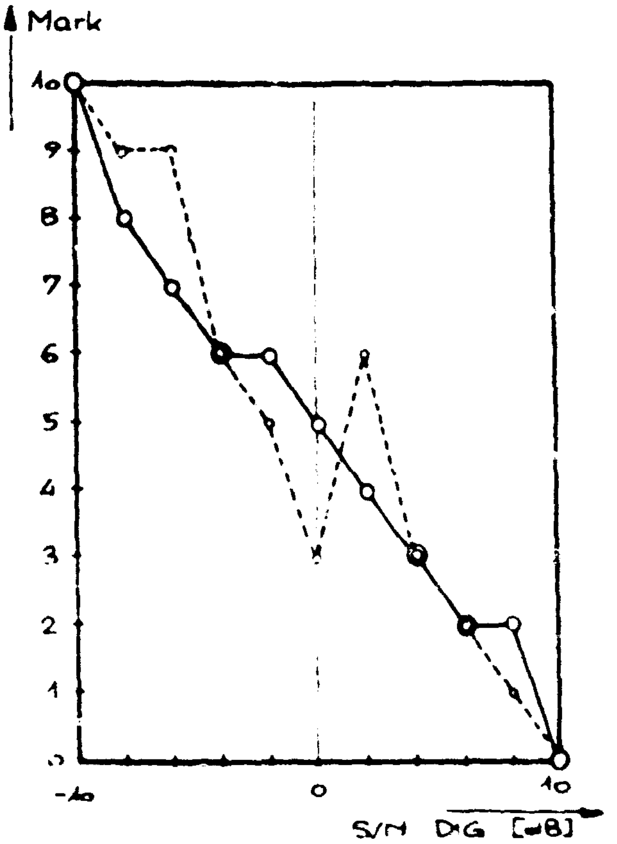


Fig. 3.20

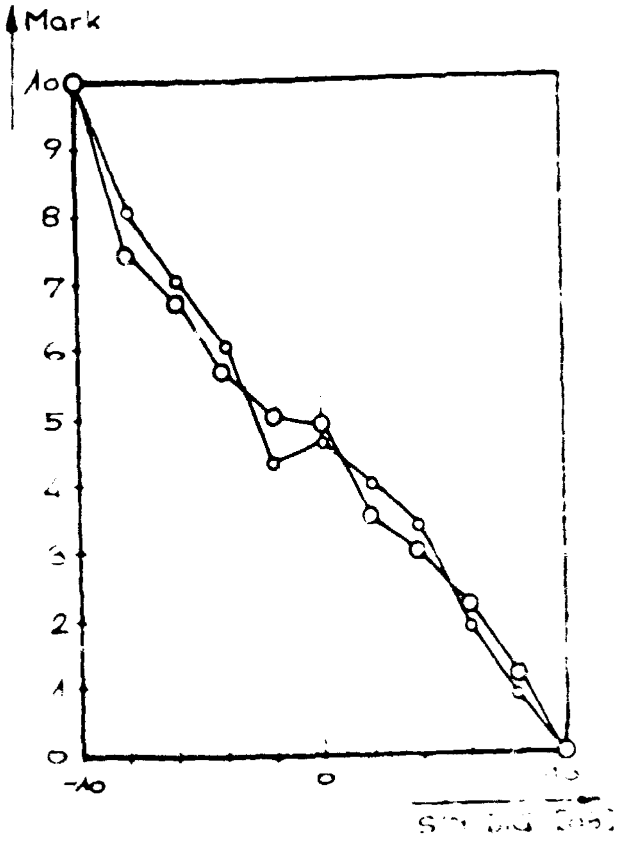


Fig. 3.21

The results of the "best" and the "worst" listener are given in Fig. 3.20. Nearly every listener has at least one or two wrong decisions. The test results of the two groups with 7 and 5 listeners are shown in Fig. 3.21.

In a third test with DIG signal and 5 listeners the incremental degradation steps have been reduced to 1 dB. The total range remained again 20 dB. Again the listeners had to use the marks one to nine for their classification. The results of this test as mean values of all

5 listeners, shown in Fig. 3.22 are very confused and indicate by no means a correct rank ordering of the 19 signals into 9 positions. By plotting the first half of the classification of the presented signals

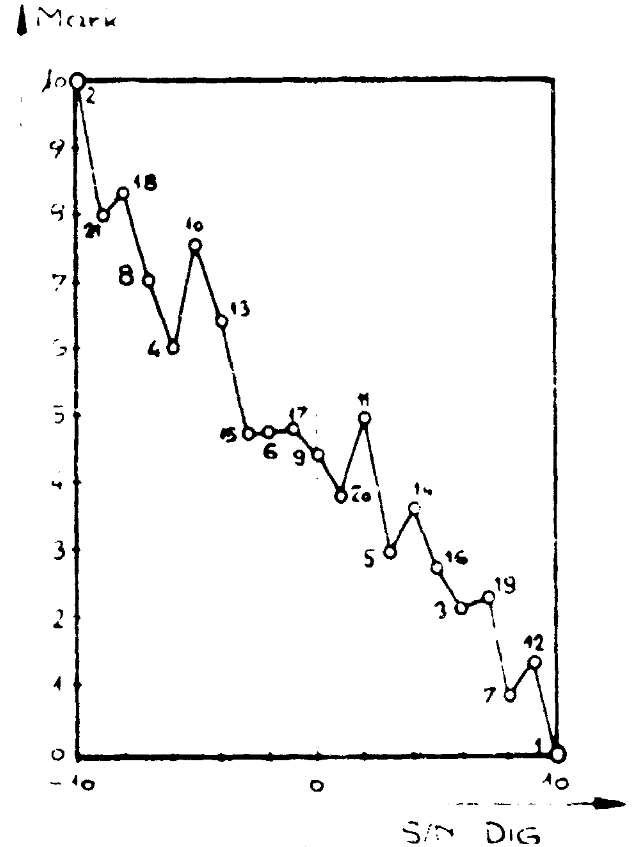


Fig. 3.22

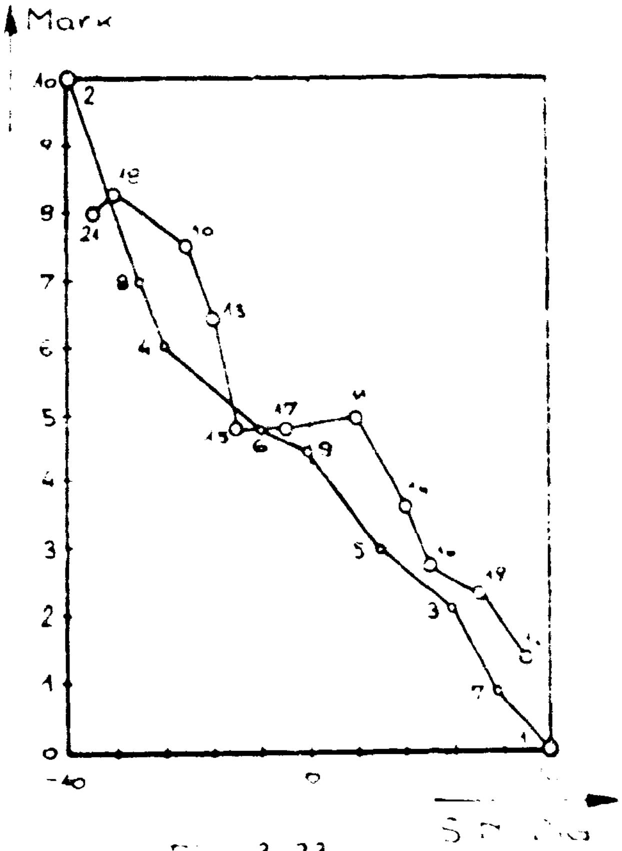


Fig. 3 23

and then the following second one, two different curves are obtained as given in Fig. 3.23. While the curve representing the first half of the decisions shows the expected monotonous decrease similar to the curves in Fig. 3.21 the second half has at least three wrong values. is not in correspondence with the first curve and is therefore the reason for the confusing shape of Fig. 3.22. The results give rise to the supposition that the listeners are able to remember the initial presented signals with the extreme qualities only during a limited number of presentations of other speech signals. If, as in the present example more than 10 different signals within this quality range are presented, or if the set of 10 signals is repeated, the listeners get more confused in their judgements.

These first tentative tests show that we are still far from an answer to the questions given above and that for the understanding of rank order tests further studies will be necessary.

Pulse delta modulation

With five speech signals generated by a pulse delta modulation system with sampling frequencies from 7,2 kc to 120 kc (PD1 - PD5 from Sec. 3.22) preference tests have been performed with a number

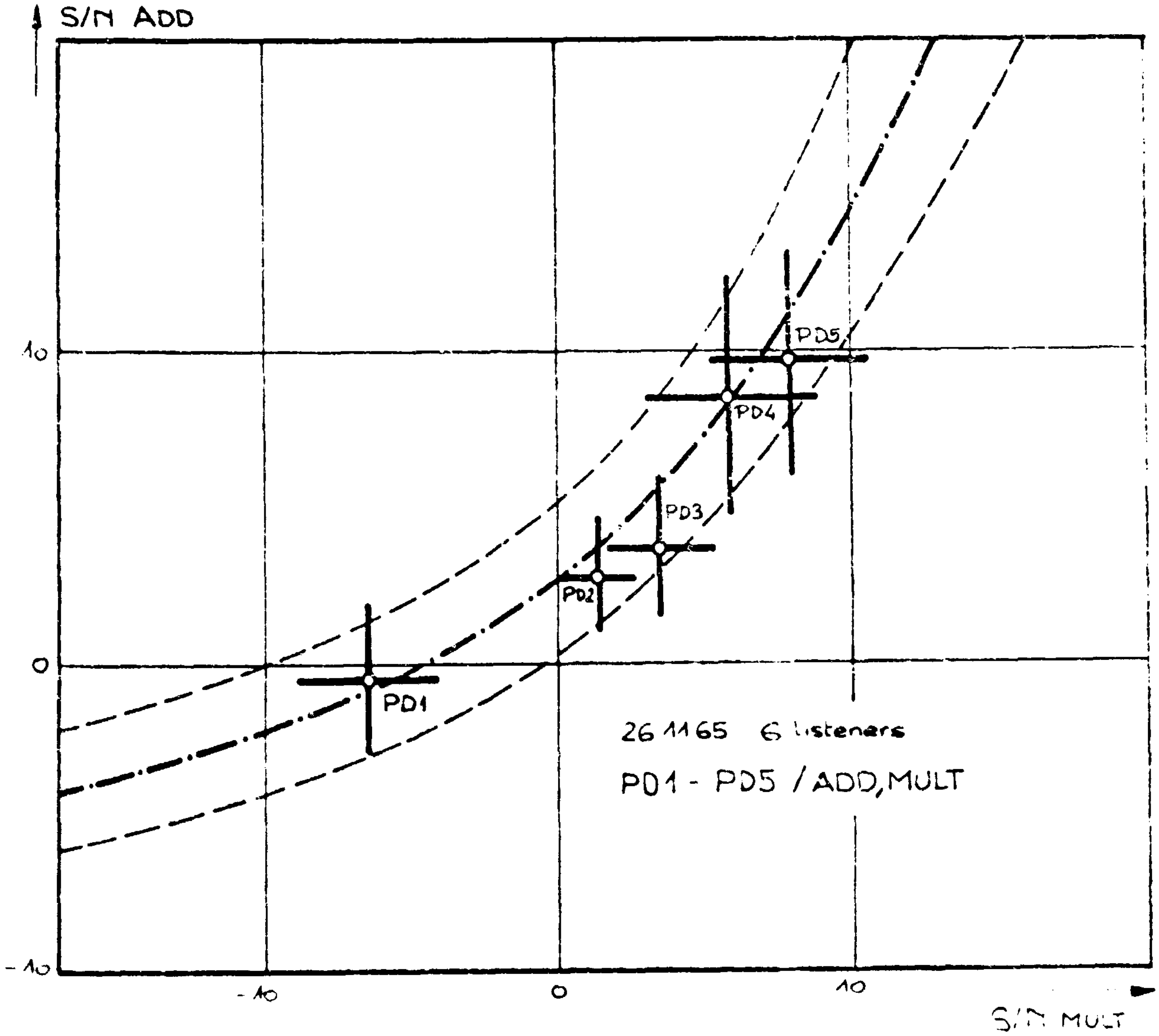


Fig. 3.24

of 6 listeners and ADD and MULT as reference signals. The mean values μ and deviations σ for the whole group are given in Fig. 3.24 for both reference signals ADD and MULT. The results show the expected rank ordering. It may be interesting to note that all the five delta modulation signals are lying not on but just beside the standard isopreference curve for ADD-MULT entirely within the uncertainty range of ± 2 dB which has been defined before.

Digital reference

Among the first informative studies with the new reference signal DIG, we made preference tests with the four standard test signals VO2, VON, LP and HP. The results from a test with 5 listeners are given in Table 3.5. The mean values M of the single listener decisions are listed and the mean values μ and corresponding standard deviations σ of the group are given in the last two columns. The mean values of the group and with it the points of isopreference are in the same order as in tests with the ADD and MULT reference. The σ-values seem to be here a little bit larger but for reliable statements further measurements are needed.

Listener Test Signal	PAZ	PRA	STE	GRZ	MAR	GROUP	
	M	M	M	M	M	μ	σ
VO2	2.5	3.5	2.5	2.5	1.5	2.5	0.8
VON	4.5	4.5	0.5	2.5	8.5	3.9	3.0
LP	4.5	7.5	5.5	4.5	10.5	6.3	3.5
HP	5.5	5.5	7.5	8.5	11.5	8.0	3.1

Table 3.5

3.12 Intelligibility

Intelligibility is a very important factor of the overall quality of speech signals and it is of special interest therefore to study all our test material also with respect to this parameter.

It has been already mentioned in Sec. 2.23 that we used monosyllabic German words from HAHLBROCKS "Freiburger Wörterteste" for our intelligibility tests. These lists are similar to the English pb-word lists. Our four main test signals have been measured with two different groups of about 8 listeners each on 6 different days. A complete test list included about 100 words. All results obtained have been averaged and are given in %-intelligibility in Table 3.6

Testsignal	HP	LP	VoN	Vo2
Intelligibility [%]	89.2	73.8	65.3	69.2

Table 3.6

In contrast to the results from preference tests where VON showed a higher isopreference level than VO2, here VO2 has an intelligibility score of about 69 % and VON only one of about 65 %. The different rank orders in view of these two aspects are an indication that preference judgements can only partially be based on estimates of speech intelligibility.

The intelligibility of the references ADD and MULT was also measured. At different S/N ratios of ADD, 400 words have been presented to 17 listeners on the same day in a random order. The results are shown in Fig. 3.25. A smooth curve may be drawn through the data points which starts at a S/N ratio of - 9 dB with a word intelligibility of about 20 % and increases monotonously.

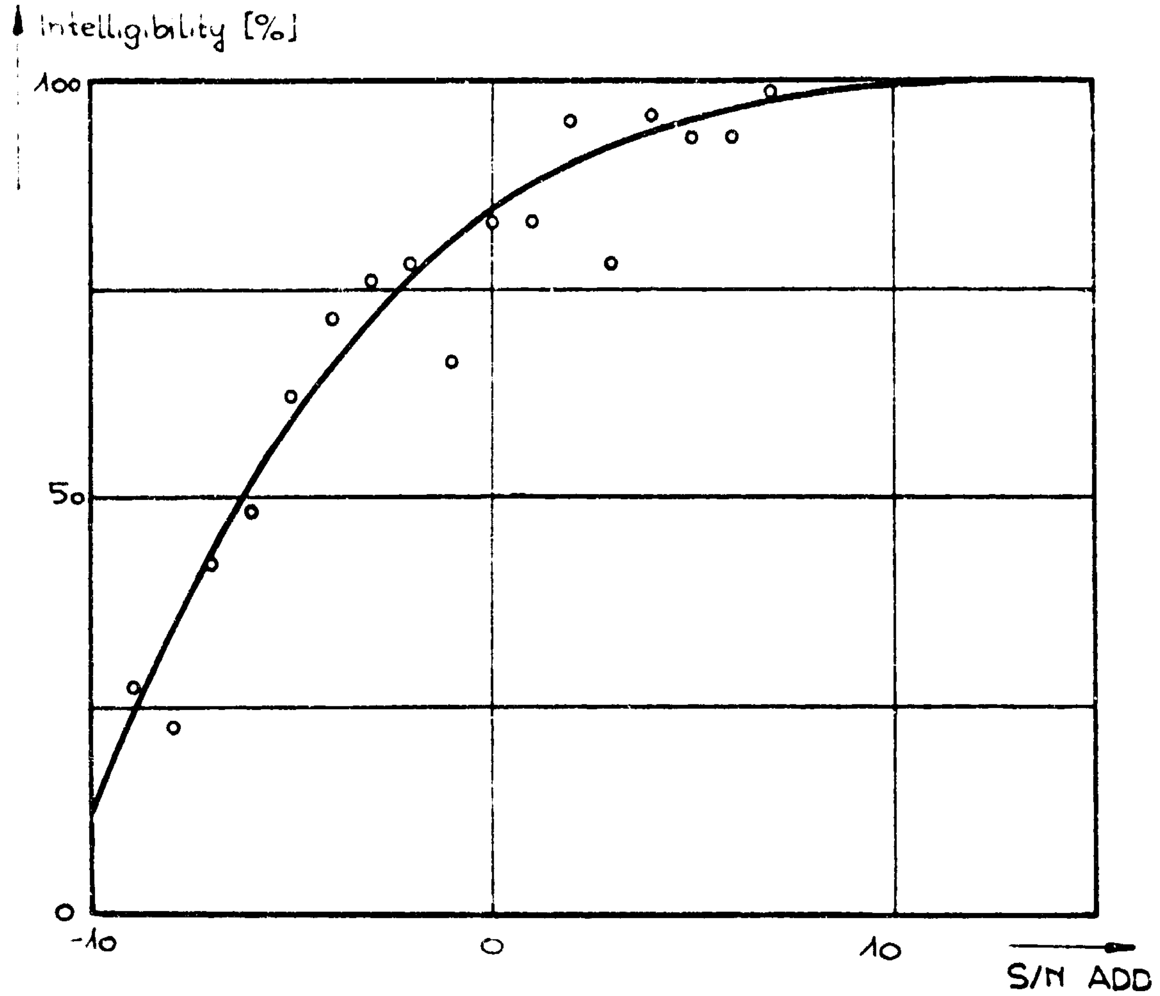


Fig. 3.25

At a S/N ratio of + 2 dB the intelligibility score reaches about 90 % and is then obviously tending to 100 % for higher values of the S/N ratio, i.e. for nearly undistorted hifi speech signals.

In the same way Fig. 3.26 shows results from 16 listeners on only one day for the signal MULT. Here the smoothed curve starts at a S/N ratio of - 12 dB with the already high value of more than 87 % word intelligibility.

That means that over the whole useful range of MULT, the signal is highly intelligible. The multiplicative distortion signal alone, i.e. hifi speech times noise was found to have 88 % word intelligibility.

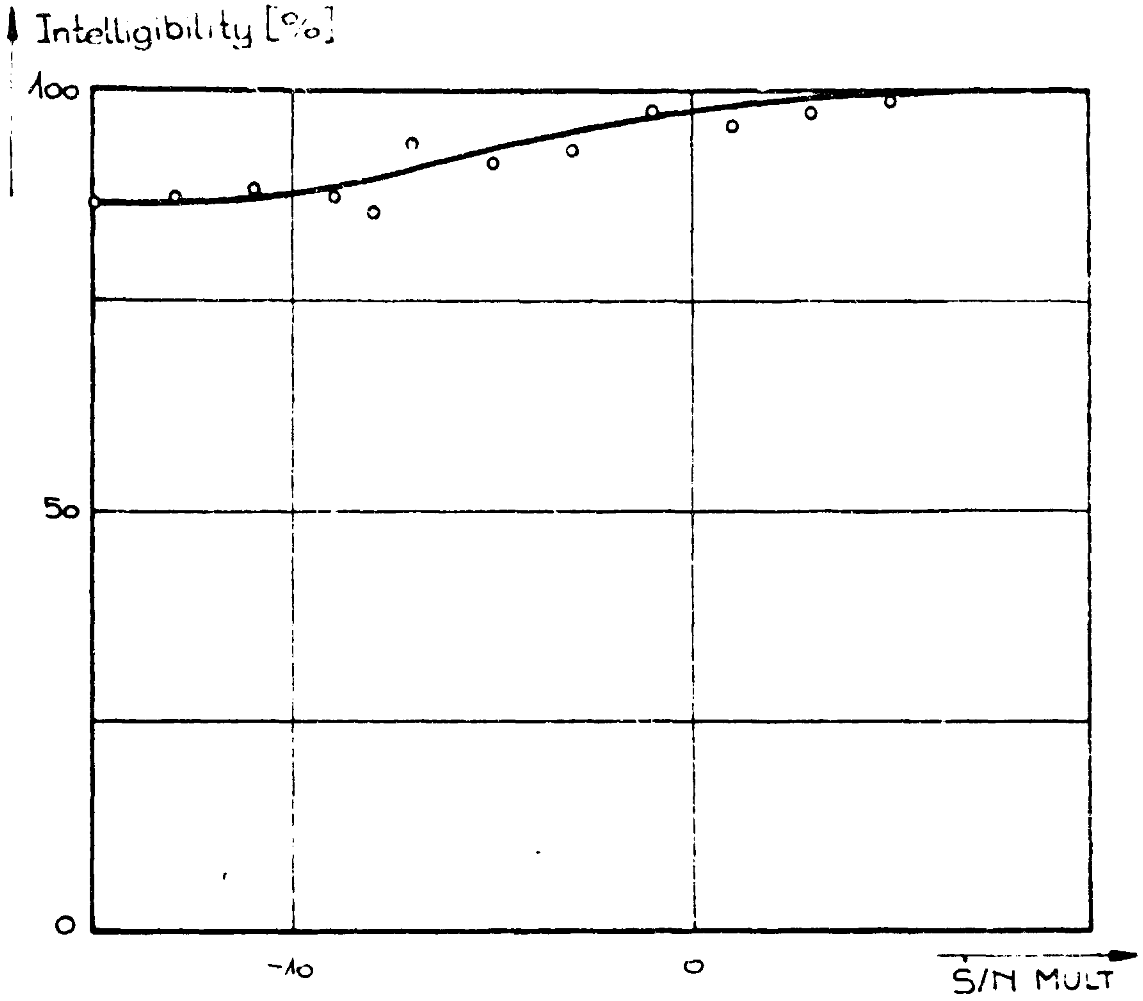


Fig. 3.26

As far as the signal DIG is concerned until now we have only a few values for the distortion signal alone. The dependence of intelligibility on the clock frequency has already been shown in Fig. 2.10. The results given in Table 3.7 repeat only that the word intelligibility is as low as about 32 % at a clock frequency ot 4 kcps.

Clock Frequency [kcps]	1	2	4
Intelligibility [%]	91.5	55	32

Table 3.7

This low value is a noticeable advantage for the purpose of degradation of a hifi speech signal compared with the high value of 88 % of the multiplicative distortion signal.

3.13 Loudness

With respect to our pair comparison prefe rence tests, there are two principal loudness measurement problems :

The determination of the optimum loudness, i.e. actually the optimum speech level for the presentation of a certain signal.

The S/N ratio of a reference signal, i.e. the loudness or level relations of a speech signal and its accompanying distortion signals.

Before these two questions can be addressed directly the intricate problem of defining and discussing of terms in this area has to be attacked.

Level definitions

In loudness and level problems we have to discriminate between :

a) Level measurements e.g. optimum speech level, or S/N ratio of reference signals

b) Level adjustments e.g. quick reproduction of a certain test condition.

The simplified blockdiagram in Fig. 3.27 of the test set-up may serve for the following explanation of the check points and the relations between the respective levels.

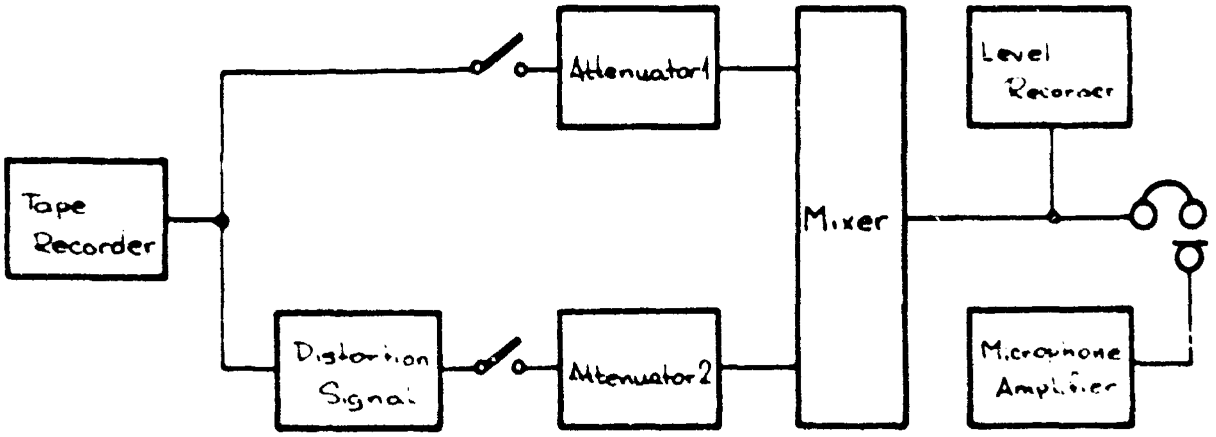


Fig. 3.27

ad a) Measurement of all sound pressure levels (SPL) produced by the earphones are done with a condenser microphone and amplifier according to the arrangement described in Sec. 2. 34. The results obtained from these measurements are given in dBA and will be referred to as speech level (SL), pilot level and distortion level for the corresponding A-weighted SPLs of speech signals, pilot tones, or distortion signals.

The subjective determination of the optimum speech level will be described in the next subsection. The measurement of the S/N ratio of reference signals depends on the nature of the respective distortion signals. The first term of the S/N ratio

$$S/N = \text{speech level} - \text{distortion level}$$

is the speech level, which can be measured only with a very low accuracy of about ± 3 dBA. The level of the additive distortion signal is a noise level and therefore measurable with sufficient accuracy. The distortion level of both multiplicative references MULT and DIG is the dBA value of the term "speech times noise". In these cases the accuracy of level measurements would be also very low. We therefore used a pilot tone recorded onto the same tape for measuring more accurately the level of a fictive distortion signal : "pilot tone times noise" instead of "speech times noise" and are now able to replace the S/N ratio for both multiplicative reference signals by

$$S/N = \text{pilot level} - \text{fictive distortion level.}$$

This ratio is of course measurable with a much higher accuracy, than the other one using the speech levels. The determination of the S/N ratio in this way is much more precise than that of the additive reference signal. This is another important advantage of the multiplicative reference signal in comparison with the additive one.

ad b) For simple and quick adjustments of any level during a test run we use a graphic level recorder connected to the transmission line from the program switch to the headsets. Level readings from this recorder are given in dBLR (level recorder) which are rms-voltage levels referring to 10 mV. Whenever possible all level measurements and adjustments of speech or distortion signals are carried out with the pilot tone and the fictive distortion signal derived from this pilot tone utilizing the level recorder. The justification for this simplification is based on the fixed relation between the pilot level (dBA) and the level of the pilot tone as a reading on the level recorder, which may be called sinus level (dBLR). This constant difference between pilot level (dBA) and sinus level (dBLR) is 74 dB, because 4 dBLR, i.e. 20mV of a 1 kcps sinus voltage, produce an A-weighted SPL of 78 dBA on the earphones.

The relation between pilot level (dBA) and speech level (dBA) is depending upon the recording conditions and therefore is different for different speech signals. Examples for our hifi speech and VON recordings are given in Table 3.8.

Signal	Sinus Level	Pilot Level	Speech Level
	dBLR	dBA	dBA
Hif.	4	78	71
Von	10	84	76

Table 3.8

During a preference test these speech levels are held constant, but in case of tests concerning optimum loudness it is necessary to vary the level of the speech signal. The speech level may be varied by attenuator 1 (Fig. 3.27). The relation between the speech level, when attenuator 1 is set to zero and the respective attenuator setting for the actually desired speech level SL is given by the equation

SL = speech level (attenuator = 0) - attenuator setting

In preference tests we have to measure S/N ratios in terms of which the reference quality is given. The S/N ratio of a reference signal

$$r(t) = s(t) + k.d(t)$$

may be defined by its levels as

S/N(level definition) = speech level(dBA) - distortion level(dBA)

The value of the constant factor k and thereby the distortion level can be changed by attenuator 2 (Fig. 3.27). Assuming that for $k = 1$ attenuator 2 is set to zero, one can derive the following equation

S/N = speech level - distortion level(attenuator = 0) + attenuator setting

Now the initial two questions of this section shall be discussed.

Optimum loudness

In our version of the isopreference method all speech signals, i.e. test or hifi signals, are presented to the listeners in a setting of optimum loudness. This loudness is determined separately for each signal in a previous subjective test. The principle of this simple loudness comparison is the same as for preference testing and consists in the alternate presentation of the same speech signal only with different speech levels (SL). The listeners now have to choose the samples within a signal pair with the preferred loudness. A simplified block diagram in Fig. 3.28 shows the test set-up. The speech signal is transmitted to the program switch over two separate channels, each containing attenuator and amplifier. The program switch generates two repeated sample pairs ABAB or BABA of the same speech signal, but with different SL. The SL of sample B is below that of sample A by a constant difference. Over a suitable loudness range

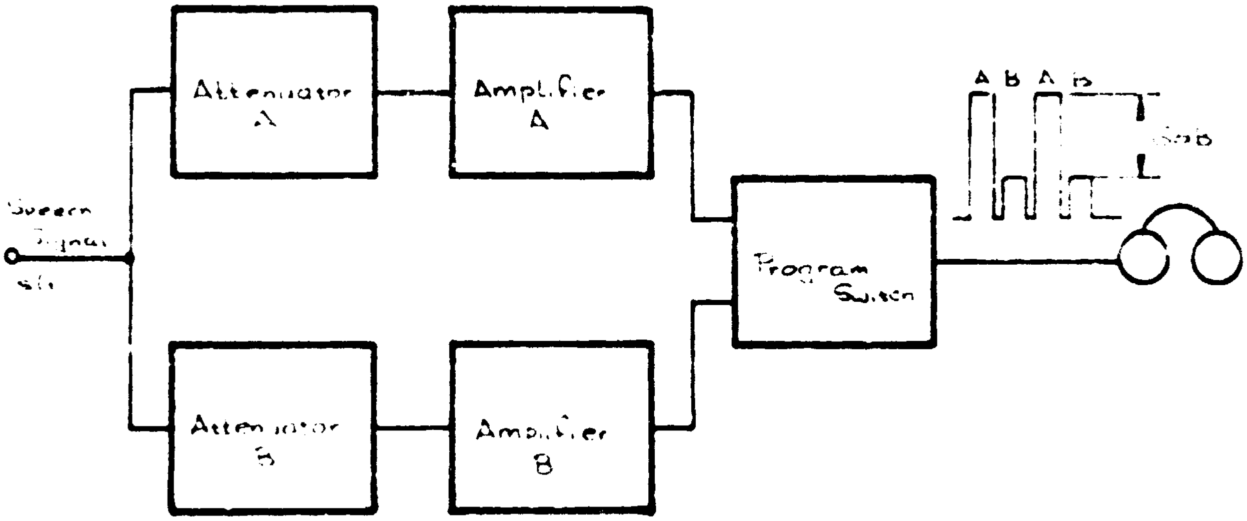


Fig. 3.28

with incremental steps of 1 dBA sample A is now vari^d in a random order. Presentation of ABAB and BABA is randomized also.

We found that 6 dB is a suitable value for the constant difference between A and B. A smaller loudness difference (4 dBA) yielded too many wrong decisions, because the listeners have difficulties to discriminate the loudness of both speech samples. A larger value (8 dBA) expands the uncertain range of the optimum loudness value and reduces therefore the accuracy of the obtained results.

The following fictive example deals with a set of errorfree data, which might have gained from an ideal listener without any "wrong" decision as results from one test run. The speech signal under test shall really have an optimum loudness which is assumed to be 70,5 dBA. Then the listener will prefer from all pairs always the speech sample with the level closer to this optimum. Ideal data of this kind are shown in Fig. 3.29

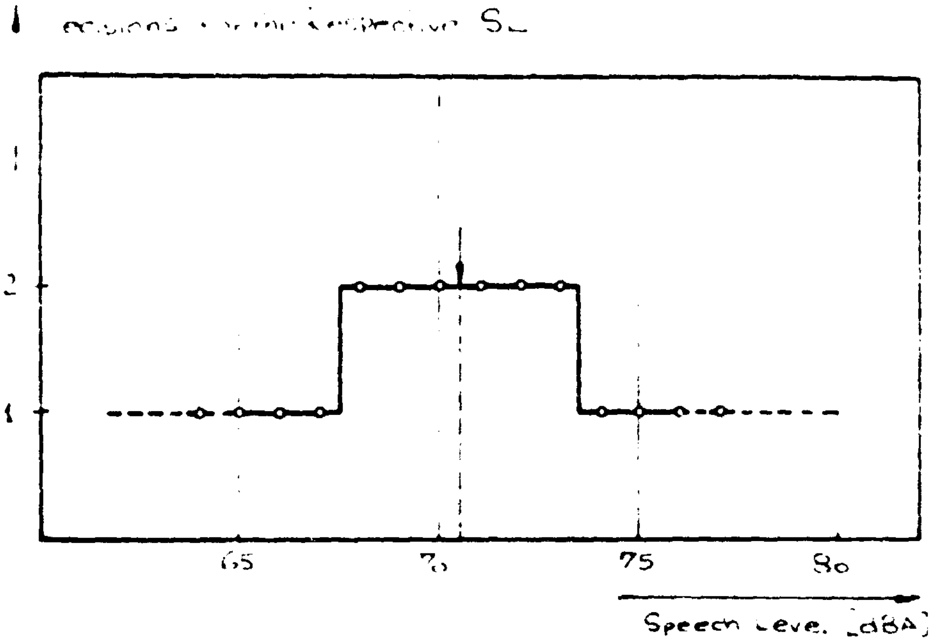


Fig. 3.29

The abscissa is numbered in SL (dBA) of the speech samples, the ordinate gives the preferences of the ideal listener at this SL.

The middle of the resulting "block" formed by the SL values which have been preferred two times, is defined now to be the optimum SL of the speech signal. Experience shows that the optimum loudness of a signal is not very critical and one should speak rather of an optimum loudness range.

Table 3.9 and Fig. 3.30 show as an example the results of a real test with 8 listeners for our hifi speech signal. All pairs were presented twice in a random order, one has therefore at least 16 decisions for each value of the SL The evaluation of this test yielded an optimum SL of 71 dBA. This SL of 71 dBA corresponds to 77 dBA pilot level and 4 dBLR sinus level as defined earlier.

Speech Level	S. A		[illegible]		[illegible]		[illegible]		[illegible]		[illegible]		[illegible]		[illegible]		[illegible]
dBA	Test 1	Test 2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	[illegible]
58	0	0	0	0	0	0	0	0	0	[illegible]	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]
59	0	0	0	0	0	0	0	0	0	0	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]
60	0	0	0	[illegible]	0	0	0	0	0	0	0	0	0	[illegible]	[illegible]	[illegible]	[illegible]
61	0	0	0	0	0	0	0	0	0	0	0	0	0	[illegible]	0	[illegible]	[illegible]
62	1	0	0	0	0	0	0	0	0	0	0	0	0	0	[illegible]	[illegible]	[illegible]
63	2	1	1	1	1	1	1	1	1	1	1	1	1	1	[illegible]	[illegible]	[illegible]
64	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	[illegible]7
65	1	1	1	1	1	2	1	1	1	1	1	1	2	1	1	1	[illegible]
66	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	[illegible]
67	1	2	[illegible]	2	[illegible]	1	2	[illegible]	1	[illegible]	1	2	1	[illegible]	1	1	[illegible]
68	2	1	1	1	[illegible]	1	1	1	1	1	1	1	1	1	1	[illegible]	[illegible]
69	1	[illegible]	1	1	[illegible]	2	2	[illegible]	[illegible]	1	[illegible]	1	2	1	[illegible]	[illegible]	[illegible]
70	1	2	2	2	2	2	1	1	1	[illegible]	2	1	[illegible]	2	[illegible]	[illegible]	[illegible]
71	2	2	2	1	2	1	[illegible]	2	[illegible]	1	2	[illegible]	1	[illegible]	1	1	[illegible]
72	1	1	2	2	2	2	2	2	1	1	2	[illegible]	1	1	1	2	[illegible]7
73	1	1	2	1	1	2	1	2	[illegible]	1	2	2	2	1	1	1	[illegible]
74	1	2	1	2	1	[illegible]	2	2	[illegible]	1	2	2	[illegible]	0	[illegible]	1	2[illegible]
75	1	1	2	1	1	1	1	1	1	[illegible]	1	1	1	2	1	2	[illegible]
76	1	1	1	1	1	1	1	1	1	1	1	2	1	1	1	1	[illegible]
77	0	0	1	1	[illegible]	0	0	0	[illegible]	1	0	0	0	1	1	1	[illegible]3
78	0	0	0	0	0	0	0	0	1	1	0	0	1	1	0	[illegible]	[illegible]3
79	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	[illegible]
80	0	0	0	[illegible]	0	0	0	0	[illegible]	0	0	0	0	1	0	[illegible]	[illegible]
81	0	0	0	1	0	0	0	0	[illegible]	0	0	0	0	0	[illegible]	1	[illegible]

Table 3.9

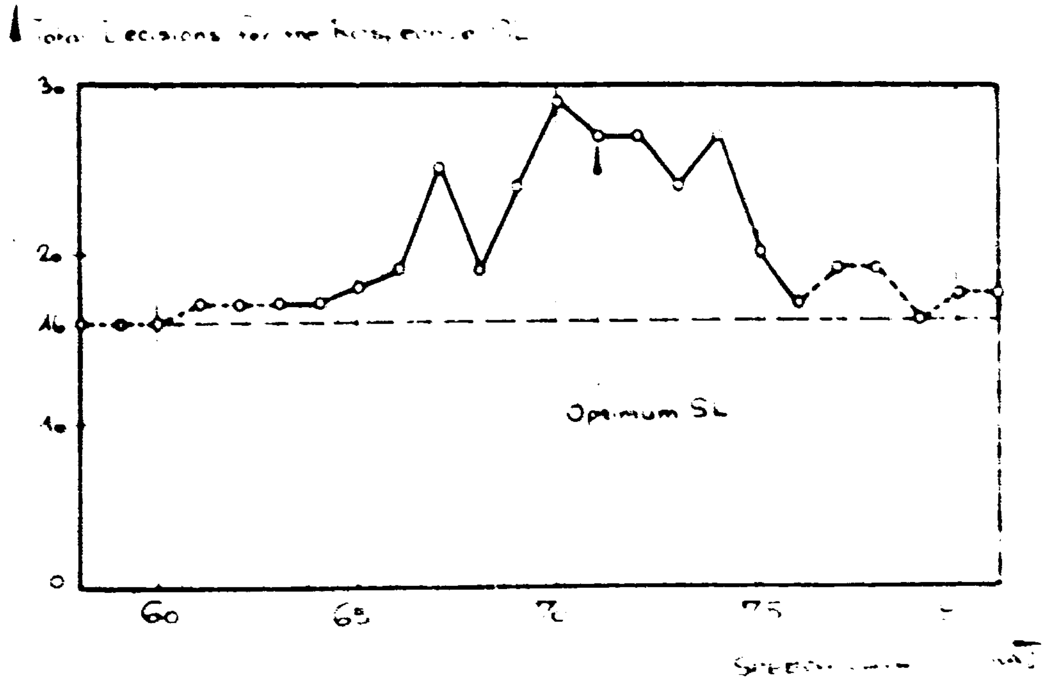


Fig. 3.30

Besides the data for the hifi speech signal, the optimum speech levels of three other systems are given in Table 3.10. Additionally the respective pilot levels and sinus levels are listed.

Signal	Speech Level dBA	Pilot Level dBA	Sinus Level dBLK
HiFi	71	78	4
VoH	76	84	10
HP	71	89	16
LP	75	84	11

Table 3.10

The problem of level definitions in normal preference tests

The idea of our preference method is to measure the quality of a speech signal in terms of the S/N ratio of a compared reference signal. At least as far as the additive reference signal is concerned, the determination of the S/N ratio is in close relation with the measurement of the loudness of speech.

As long as there exists neither a close definition nor an accurate measuring procedure for the speech level, one has at least two obvious possibilities to define the S/N ratio. First, one may use, e.g. the A-weighted SPL of both speech and distortion signal, where the noise level is exactly measurable, while the speech level has to be specially defined and may be measured, e.g. as we postulate it in Sec. 2.34. The other possibility is to compare speech and noise signals directly in a subjective loudness test and to define the S/N ratio to be zero, when speech and noise signal are judged to be equal in loudness. Our experiments described in this subsection shall illustrate the relation between speech and corresponding distortion signal, i.e. between both of the S/N ratios mentioned above. The results shall show whether or not a subjective determination of the

S/N ratio is useful. As long as there is no better definition or measuring procedure proposed, it might be adopted during our studies in order to have an operational definition of the S/N ratio of a distorted speech signal.

We made experiments to compare the speech signal with its corresponding distortion signal in a loudness pair comparison test, similar to the method described above. During one test run, speech signal A is held constant at a certain level, while distortion signal B is presented with different levels in a random order. The covered range is about 12 dBA in steps of 1 dBA. A simplified block diagram of the test set-up is given in Fig. 3.31.

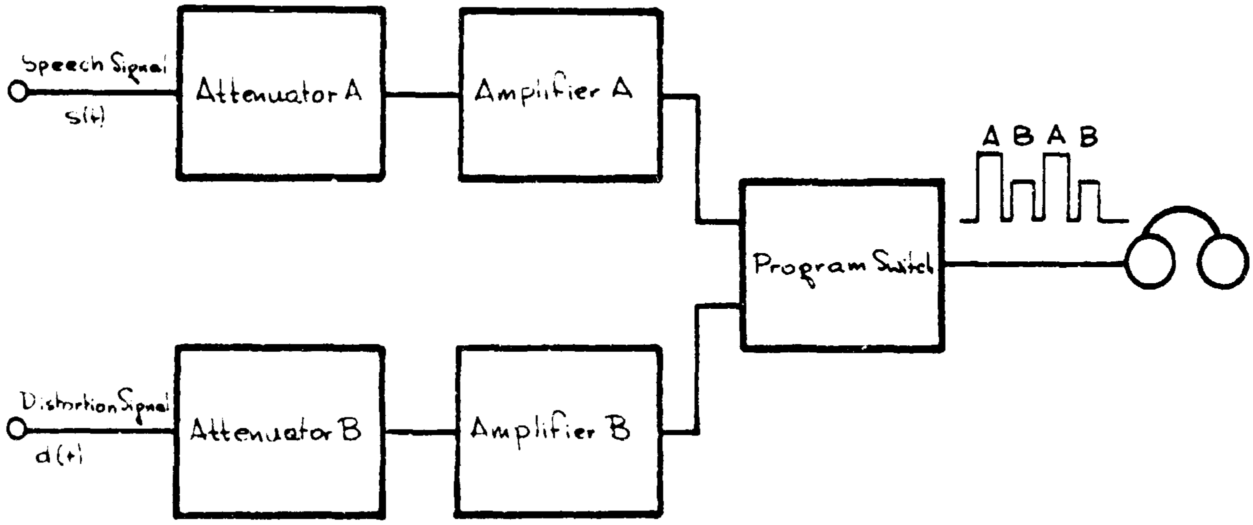


Fig. 3.31

Speech signal and distortion signal are sampled by the program switch in the usual ABAB series and presented over headphones. The listeners are requested to determine from each repeated pair the signal with the greater loudness. In this way one gets, similar as in a preference test the point of isopreference the one level of the distortion signal, which is found to be equal in loudness with the respective speech level. Thus one may express the speech level in terms of the measurable level of the distortion signal which has been judged to be equally loud.

The results of hifi speech signal compared with pink noise shall be given in Fig. 3.32 as an example for such an "isoloudness test". The measurements were performed at different dates and with different numbers of listeners as stated in the diagram. Each point represents the mean value of all listeners together with the corresponding standard deviations

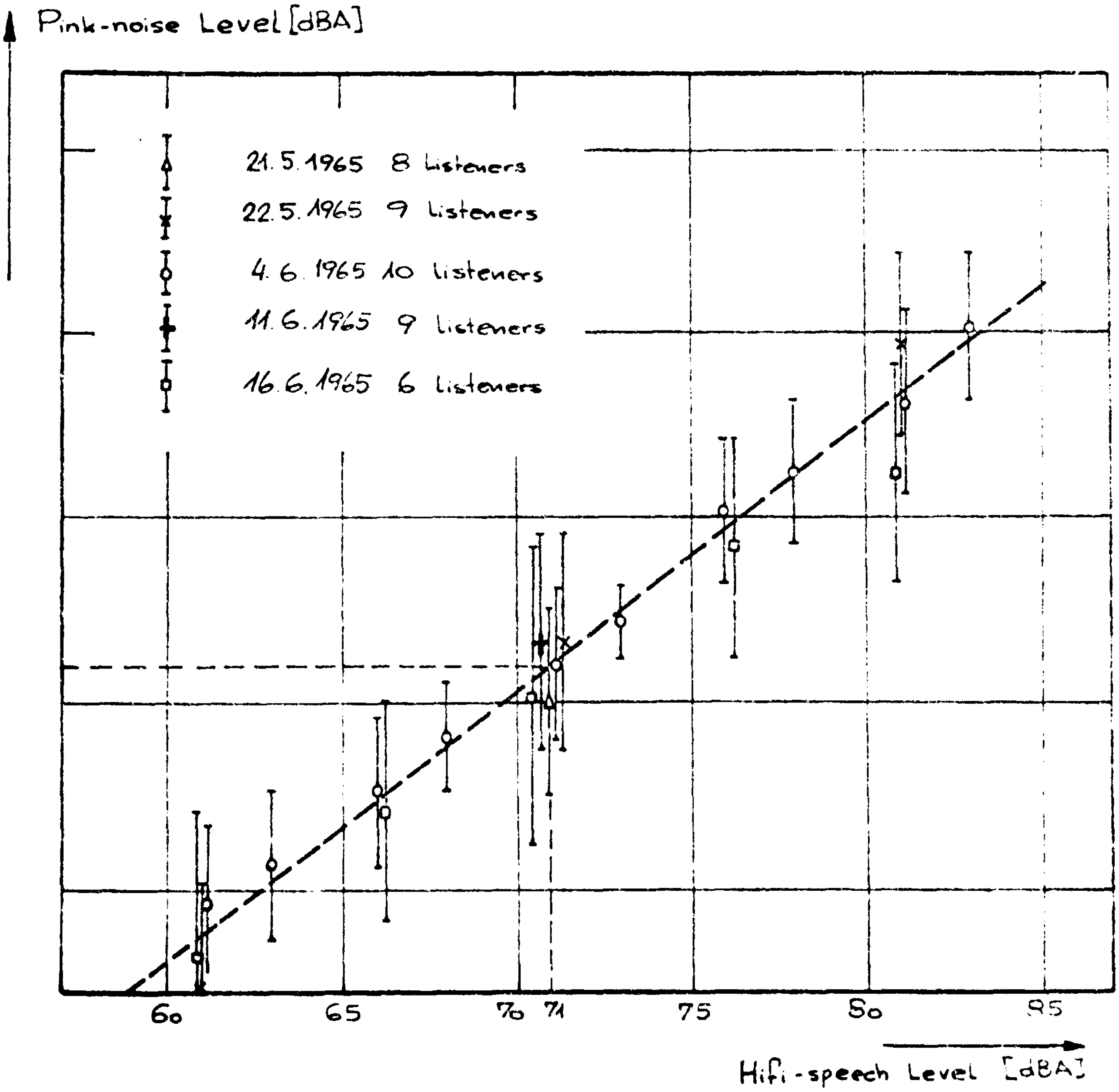


Fig. 3.32

The reproducibility of the results is obviously quite well, the values of σ are entirely within the range of the usual deviations which have to be expected in subjective measurements.

Fig. 3.32 shows an approximately linear relation between hifi speech and pink noise. The optimum hifi speech level of 71 dBA corresponds to a pink noise level of about 66 dBA. One can see that for the additive reference signal presented with optimum hifi speech level, the difference between the two definitions of the S/N ratio given above is approximately 5 dB or written as an equation :

ADD : S/N(loudness definition) = S/N(speech level definition) - 5 dB

Similar measurements carried out for the multiplicative distortion signal yielded a corresponding value of 74 dBA judged as equal loud to 71 dBA hifi speech level or in form of an equation

MULT : S/N(loudness definition) = S/N(speech level definition) + 3 dB

The corresponding values for the digital distortion signal yield the relation : 71 dBA hifi speech level is judged as equal loud in comparison with 75 dBA or

DIG : S/N(loudness definition) = S/N(speech level definition) + 4 dB

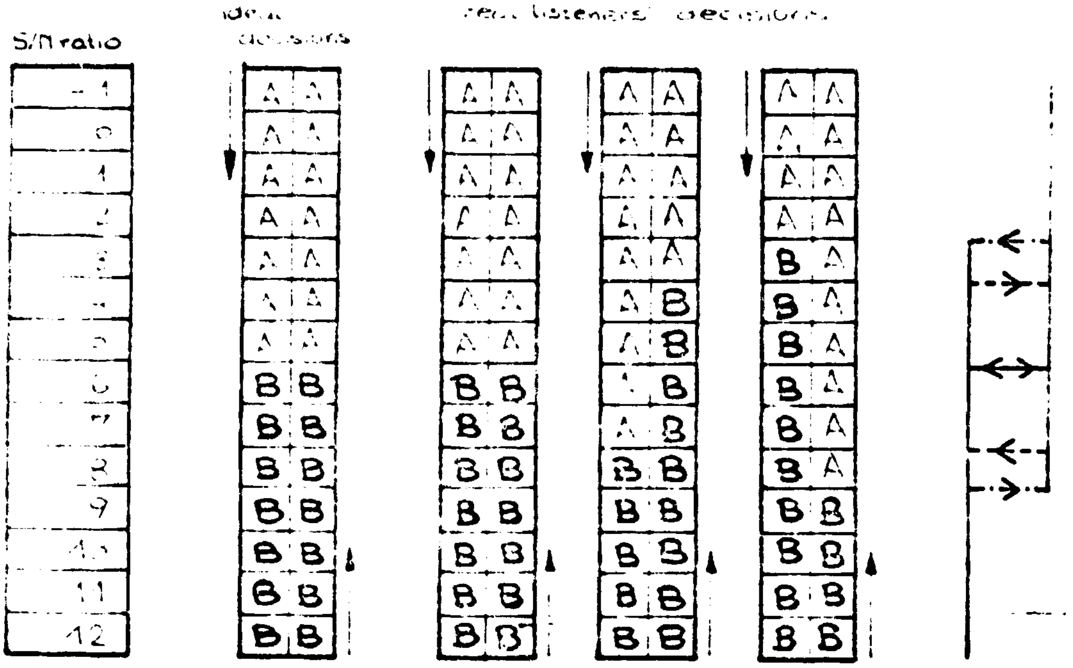
3.14 Human Factors

Obviously the main difficulty when running subjective tests with a group of listeners is the influence of human factors, most of which may differ from listener to listener and are therefore hard to identify and to consider. Several factors such as momentary personal condition, disposition or motivation may be statistically different for single listeners and will be therefore eliminated to some extent by testing a group of listeners and forming the mean values of their decisions. Other factors, e.g. hysteresis, check-on-order-effect, fatigue, or training may be common to the group. If possible these factors have to be studied separately and their influence on the test results has to be taken into consideration.

Hysteresis and randomization

To determine a point of isopreference, a listener compares a fixed test signal with a reference signal the quality of which is varied. The question arises how the reference quality in a range about the point of isopreference should be varied. Suppose, one would improve the reference quality step by step from a value which is certainly worse than the isopreference quality, to a value which is certainly better than this quality. Then up to the isopreference level an ideal listener would have only decisions which favor the test signal and beyond this point only decisions for the reference signal. Corresponding results would be obtained if the test sequence were reversed from good to bad reference quality. The listeners' decisions then would first favor the reference signal and below the isopreference level favor the test signal.

A real test with monotonously increasing (↓) and then decreasing (↑) reference quality yields different results. The results from an ideal listener in comparison with the decisions of real listeners are shown in Fig. 3.33. Being confronted with pairs consisting of test signal B and monotonously varied reference signal A, the listener notices at a certain moment that he has to change his decisions from A to B or from B to A and then he remains with the new decision. Disregarding the case of "ideal" decisions, the point where the listener changes to the other signal will be different, when presenting a row of pairs with monotonously decreasing and increasing reference qualities. That means the decisions of the real single listener show a certain hysteresis which may have maximum values of 4 - 6 dB and average



S/N ratio	ideal decisions		real listeners' decisions					
-1	A	A	A	A	A	A	A	A
0	A	A	A	A	A	A	A	A
1	A	A	A	A	A	A	A	A
2	A	A	A	A	A	A	A	A
3	A	A	A	A	A	A	B	A
4	A	A	A	A	A	B	B	A
5	A	A	A	A	A	B	B	A
6	B	B	B	B	A	B	B	A
7	B	B	B	B	A	B	B	A
8	B	B	B	B	B	B	B	A
9	B	B	B	B	B	B	B	B
10	B	B	B	B	B	B	B	B
11	B	B	B	B	B	B	B	B
12	B	B	B	B	B	B	B	B

Fig. 3.33

values of 2 - 3 dB. Of course the hysteresis may degenerate to zero, when the listener has a chain of "ideal" errorfree decisions. The hysteresis may be explained as a range of uncertainty and may be caused by an accomodation during a single test series, or by lack of attention of the listener. The hysteresis effect can be avoided by a randomization of the test material.

Additionally studies were made on the influence of preceding decisions, when presenting the reference qualities randomly. Decisions in the critical range around the isopreference level have been marked whether the preceding reference signal was of very bad quality or of very good quality. In six different tests we counted the decisions in the critical range in dependence of the quality of the preceding reference signal. The results show that presenting the reference signal in a random order to the listeners is sufficient to make the decisions practically independent from the preceding values of the reference quality

Check-on-order effect

The mode of presentation of the test material has been discussed in Sec. 2.1. Here the question shall be examined whether differences in the results have to be expected when the repeated signal pairs ABAB (abbreviated AB) are reversed to BABA (abbreviated BA). Systematic differences would indicate that there exists a "check-on-order" effect which means that the listeners have the clear tendency to be more influenced by the last speech sample.

Two different cases could be discriminated. In one case there are negligible differences between AB and BA results, in the other case we found differences which are by no means negligible, but have to be considered or may perhaps be avoided by randomizing not only the reference quality but also the sequence of presentation.

The first case shall be illustrated by a sequence of normal preference tests with 6 listeners on one day with VON and HP as test signals and ADD and MULT as reference signals. In a first test run the mode of presentation has been AB and in a second test run BA. The results of these presentations of both test sequences AB and BA are given in Table 3.11. Here we can see that the

Reference	MULT				ADD			
Presentation	AB		BA		AB		BA	
Test Signal	μ	σ	μ	σ	μ	σ	μ	σ
Von	1.7	2.4	1.5	33	5.1	3.1	5.9	29
HP	8.4	2.8	8.6	26	9.2	3.4	9.8	35

Table 3.11

differences are indeed very small. An explanation for this may be that in this type of preference test the test signal is held constant. During a test run the listeners seem to form a concept of the constant quality of the test signal in their mind. With this concept, it makes nearly no difference whether this test signal is the first speech sample or the second one of a pair. Therefore the test results are nearly the same in the two cases AB and BA.

The second case, a pronounced check-on-order effect was found to be possible in area tests between the reference signals, e.g. ADD and MULT. This test is described in detail in a previous subsection. In a first test run we presented always the AB sequence MULT-ADD and in a second run the sequence ADD-MULT with random variations of both speech qualities. The results of two different test sessions on two days with 5 to 7 listeners have been averaged and are shown in Fig. 3.34. The results from AB and BA presentation differ in a systematic manner. The first isopreference relation MULT-ADD corresponds quite good with the standard isopreference curve. The other data are still lying within the previously given uncertainty range, but have in the special case of the ADD-MULT relation a nearly constant difference of about 2 dB in one direction. Further examples with random presentations of AB and BA pairs are given in the Appendix. Both isopreference relations ADD-DIG (and reversely DIG-ADD) and MULT-DIG (DIG-MULT) show the same tendency of favoring the speech sample presented finally. The differences are not constant as in the case of ADD-MULT, but may increase to values of about 4 - 5 dB.

The examples show that the check-on-order effect may affect the test results, if the speech samples are not presented in a random sequence. Our explanation is that in area tests there is no constant test signal because both speech signals are varied randomly in quality. The listeners have no chance to form a stationary concept of the test speech quality and are then actually favoring the last presented speech sample.

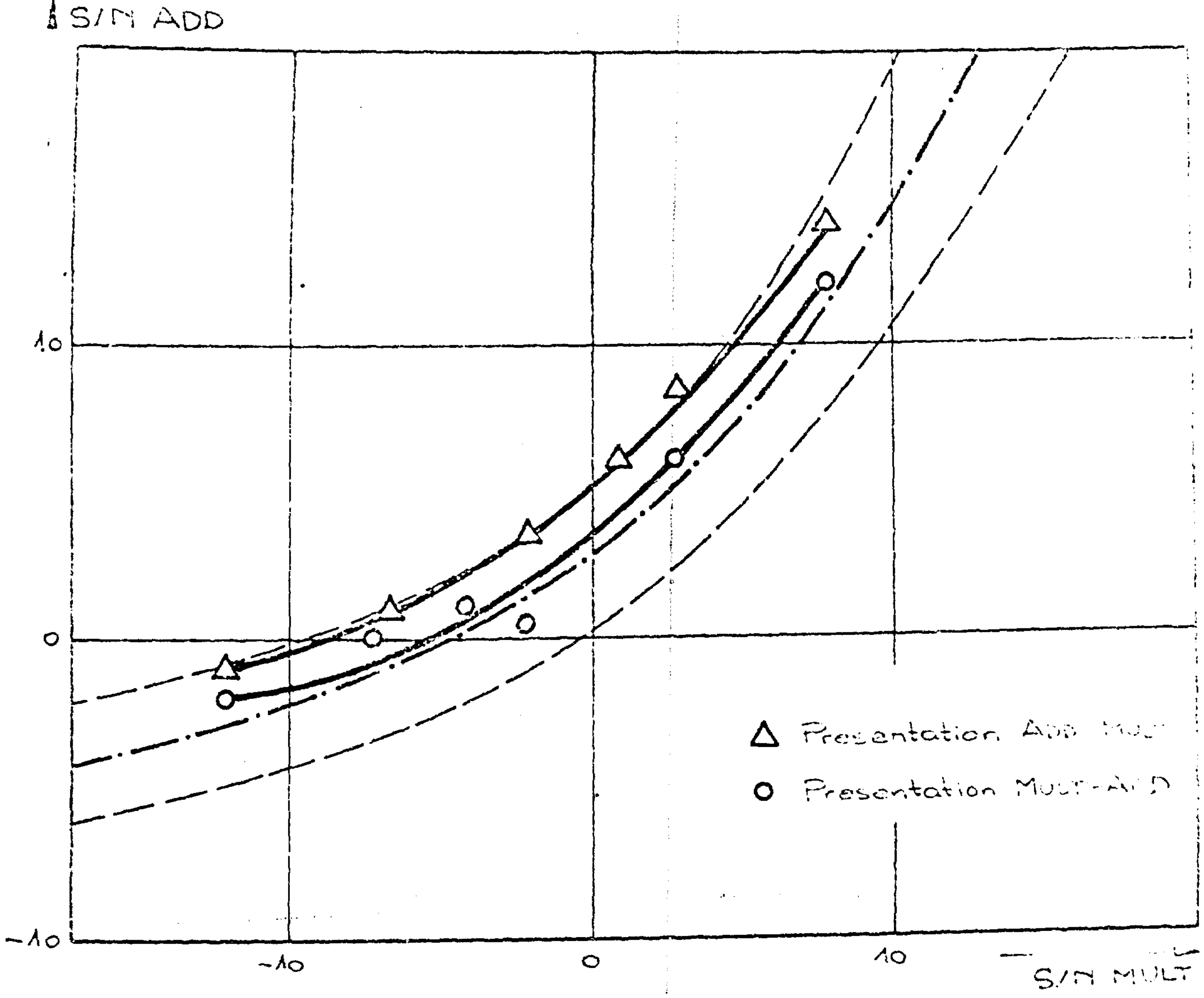


Fig. 3.34

Fatigue

We are not able to identify a systematic variation of test results from the beginning to the end of a session when considering equal tests. The scattering of the results and the reproducibility of a certain test is quite the same if the test is repeated within a test session or on the following day. After a usual test session of about 2 hours the listeners show no noticeable fatigue which might cause additional variations of the test results.

Variability of single listeners

The dependence on time of the measurement results obtained is a further question concerning human factors. Reproducible results are an important requirement for the reliability of a test procedure. The reproducibility is influenced by several subjective factors like training and learning, accomodation to a system or to a certain test signal. Of course there is also an influence caused by the inconstancy of the decisions of a single listener due to personal conditions which may have varied. But this last factor cannot be taken into account for it is different from one listener to another and has to be eliminated by averaging the results of a listener group.

Most subjective measurements require a certain training of the testing group, so that the listeners get well experienced with the test procedure. Our studies have shown that after a few preference test runs, the listeners are familiar with the test procedure. Even the first results of preference tests yield unambiguous decisions for most of the listeners, i.e. decisions preferring test signal B and reference signal A are clearly separated without any noticeable uncertainty range. Such results may be found already in the very first test run. They are considered to be a striking evidence for the certainty of the single listener in making his decisions.

Many of the listeners make again unambiguous decisions when a test is repeated on the following days and weeks. But the isopreference level for single listeners as well as for the whole group may be somewhat different compared with those of previous test sessions. The listeners are then again quite sure in making their decisions but they must have changed their minds about one or both of the presented speech signals.

Considering the theoretical case that these variations are monotonous then we may discriminate different possibilities, i.e. whether the speech signals are judged to be better, worse, or equal compared with the previous test session (Fig. 3.35).

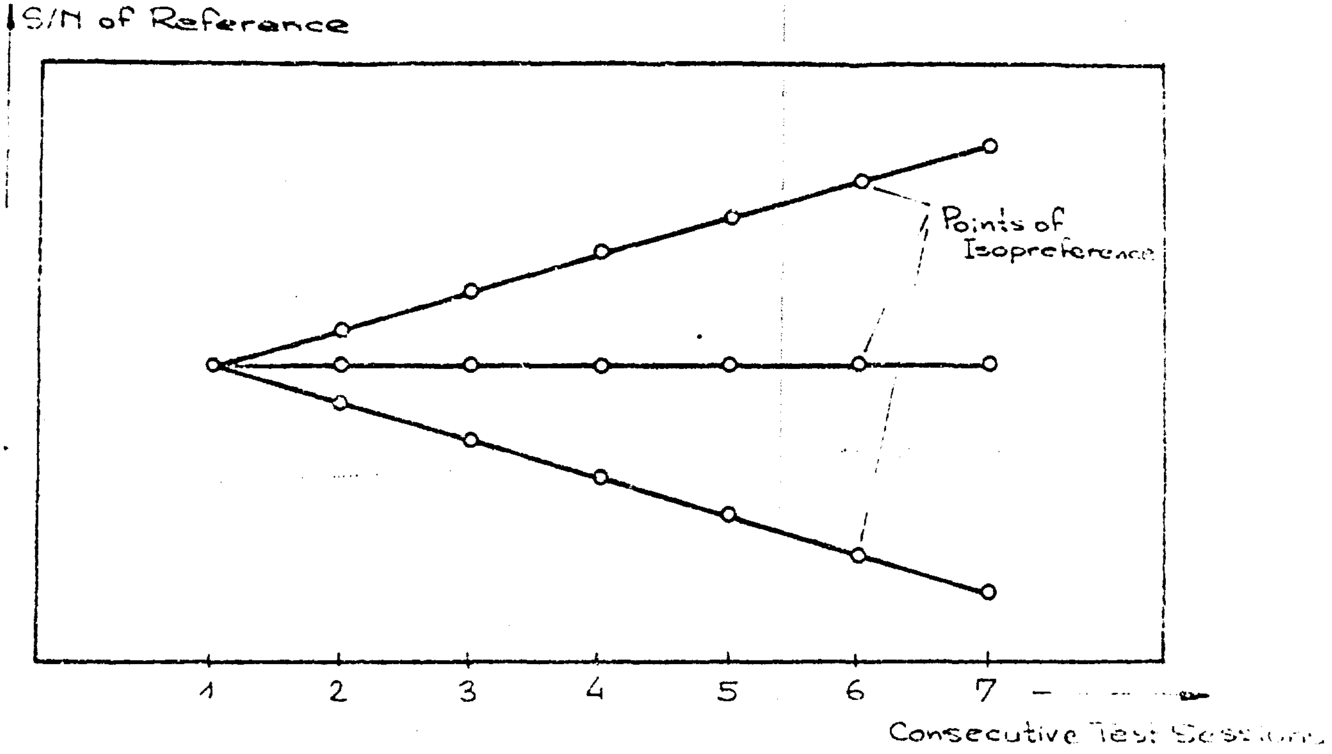


Fig. 3.35

If the isopreference level increases, this may be due to better acceptation of the test signal or to higher annoyance created by the reference signal in subsequent test sessions. It is still an open question whether it can be decided which kind of such accomodation effects occurs.

In order to get evidence for a study of the long time variability of our listeners we have frequently repeated preference tests with the four standard test signals and the reference signals ADD and MULT. The results given in Figs. 3.36 - 3.39 cover a period of about two months. The abscissa is numbered in test sessions which are approximately one week apart. The results of each reference signal with the four test signals are given in a separate diagram.

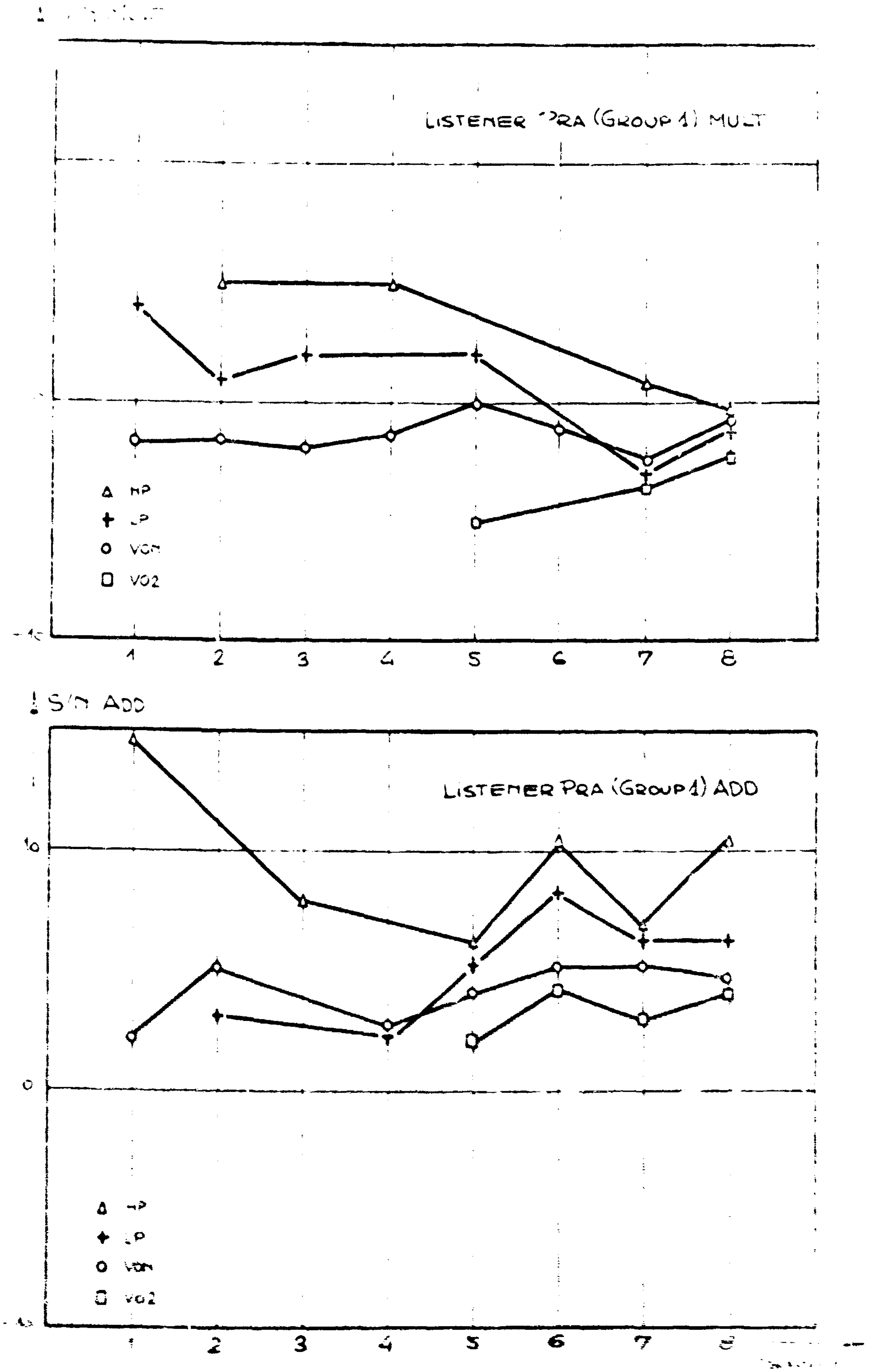

LISTENER PRA (GROUP 1) MULT
HP
LP
VOM
VO2
1
2
3
4
5
6
7
8
S/N ADD
LISTENER PRA (GROUP 1) ADD
10
0
HP
LP
VOM
VO2
1
2
3
4
5
6
7
8

Fig. 3.36

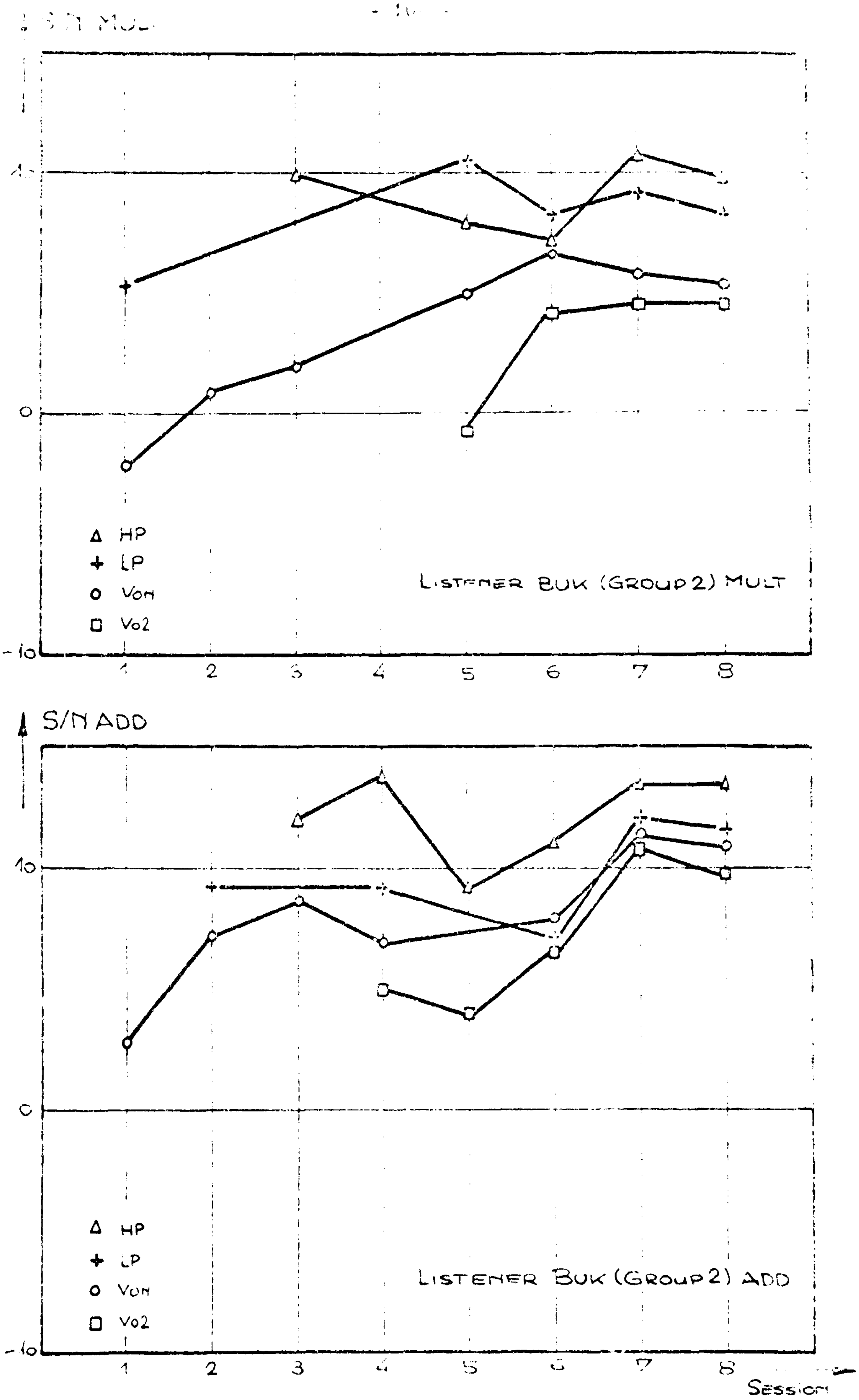

S/N MULT
10
0
-10
Δ HP
+ LP
o VOH
□ VO2
LISTENER BUK (GROUP 2) MULT
1
2
3
4
5
6
7
8
S/N ADD
10
0
-10
Δ HP
+ LP
o VOH
□ VO2
LISTENER BUK (GROUP 2) ADD
1
2
3
4
5
6
7
8
SESSION

Fig. 3.37

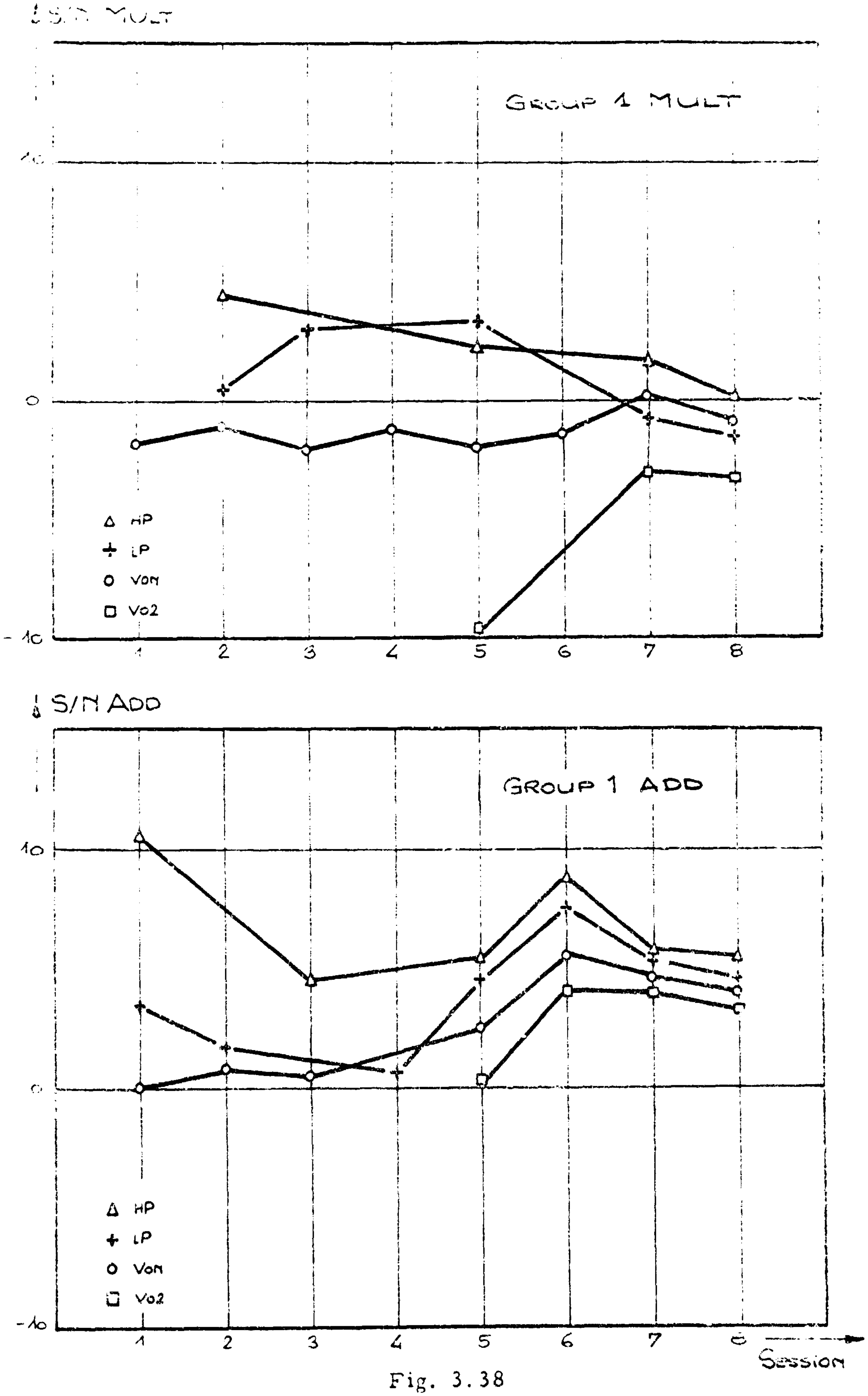

S/N MULT
GROUP 1 MULT
10
0
-10
Δ HP
+ LP
O VON
□ VO2
1
2
3
4
5
6
7
8
S/N ADD
GROUP 1 ADD
10
0
-10
Δ HP
+ LP
O VON
□ VO2
1
2
3
4
5
6
7
8
SESSION

Fig. 3.38

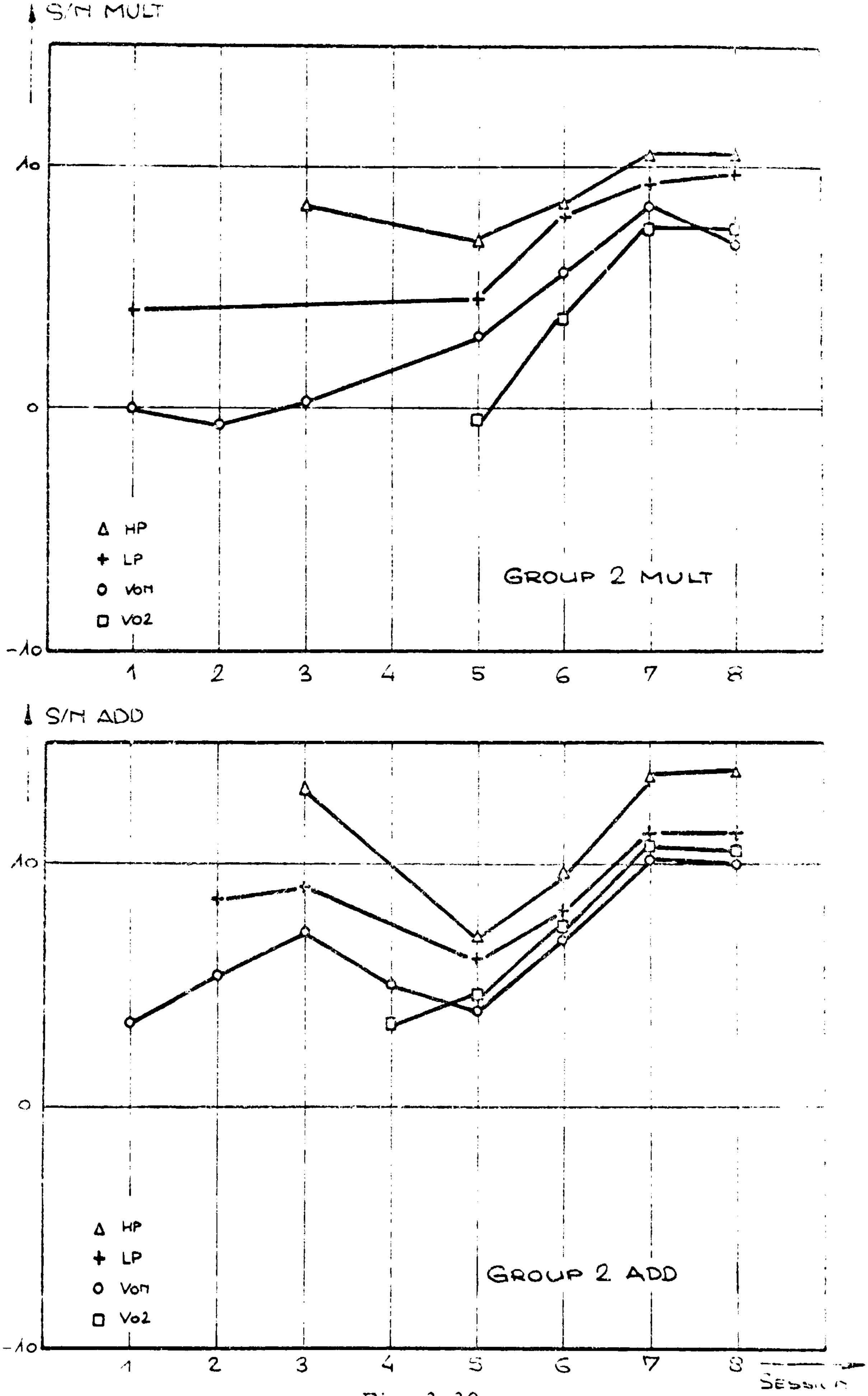

S/N MULT
10
0
-10
Δ HP
+ LP
O VoM
□ Vo2
GROUP 2 MULT
1
2
3
4
5
6
7
8
S/N ADD
10
0
-10
Δ HP
+ LP
O VoM
□ Vo2
GROUP 2 ADD
1
2
3
4
5
6
7
8
Session

Fig. 3.39

The examples show that the establishing of confidence criteria for single listeners and for groups as function of time and training is much more complex than anticipated. The accomodation to the test material is not the same for the different speech signals. This is true for the single listeners as well as for the group. Testing group 2 shows mainly increasing curves while group 1 has a slight tendency to have lower isopreference levels after a period of two months. There is only one special characteristic common to all results of single listeners and listener groups which is demonstrated in Fig. 3.40. At the beginning of the tests the four signals have been found to be clearly

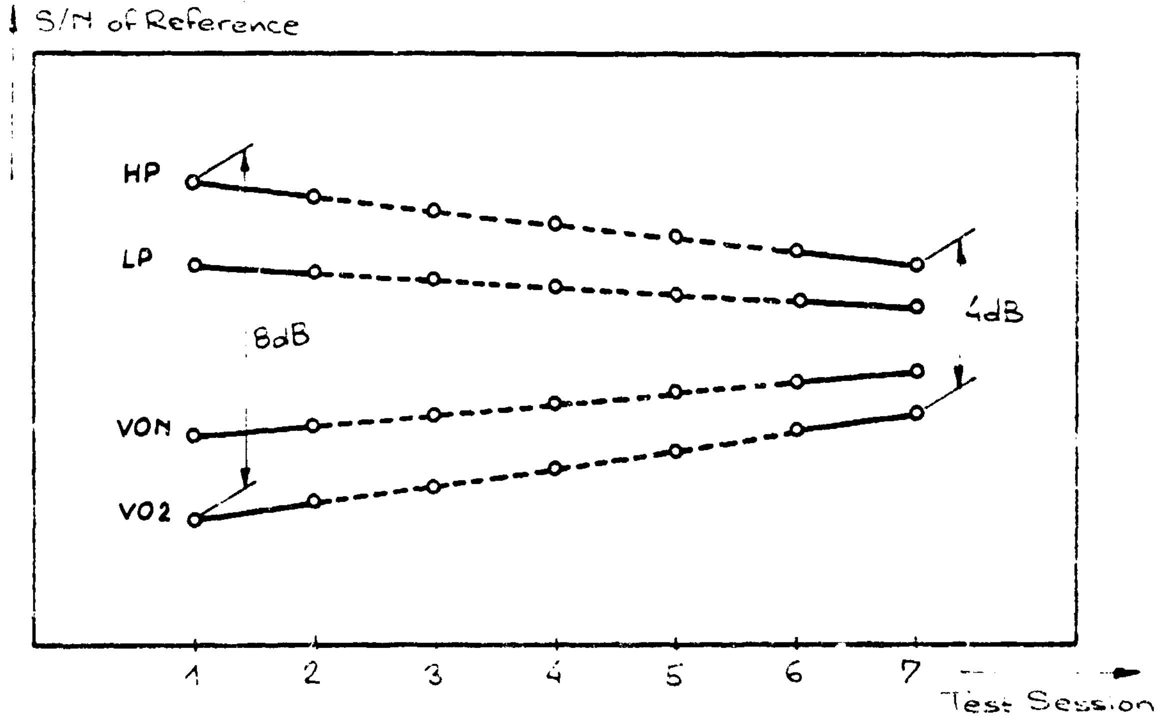


Fig. 3.40

discriminable. HP and VO2 systems are in terms of the reference system 6 - 10 dB apart. Over a period of 2 - 3 months about 30 tests have been executed. Now HP and VO2 system are only 3 - 4 dB apart. An explanation of this effect is that the listeners get more and more accustomed to the frequently presented test material. This reduction of the isopreference level differences seems to be an effect of overtraining. It is interesting to see that in spite of the reduced differences the rank order of the signals is mostly preserved.

3.2 Discussion of Preliminary Results

In the previous Section 3.1 several aspects of our study have been discussed in some detail. Now, in this section we want to summarize the main positive results and also the main difficulties we had encountered in the past research period. The section is divided into three parts : transitivity, training, and intelligibility versus preference. The first part is positive and affirmative, the other two parts reveal problems which will necessitate further work before final statements are possible.

Transitivity

Our study shall establish a reliable procedure for rating speech signals in view of their quality in terms of the hypothetical one-dimensional scale for the parameter "preference". Transitivity is a basic requirement for the existence of a one-dimensional rating scale. Transitivity exists if for any signal pair A and B as well as for the signal pair B and C which have both been found to be isopreferent, the signal pair A and C is also found to be isopreferent.

A first general test of transitivity is possible using the standard isopreference curve for ADD-MULT together with the isopreference levels for four standard test signals expressed in terms of the reference signals ADD and MULT. Fig. 3.41 shows the isopreference points together with the respective δ values of the test signals HP, LP, VON and VO2 in an ADD versus MULT diagram for a trained group. A transitivity check is given by the distances between these points and the standard isopreference curve plotted into the same diagram. All four points lie,as it should be,within the postulated uncertainty range of ± 2 dB.

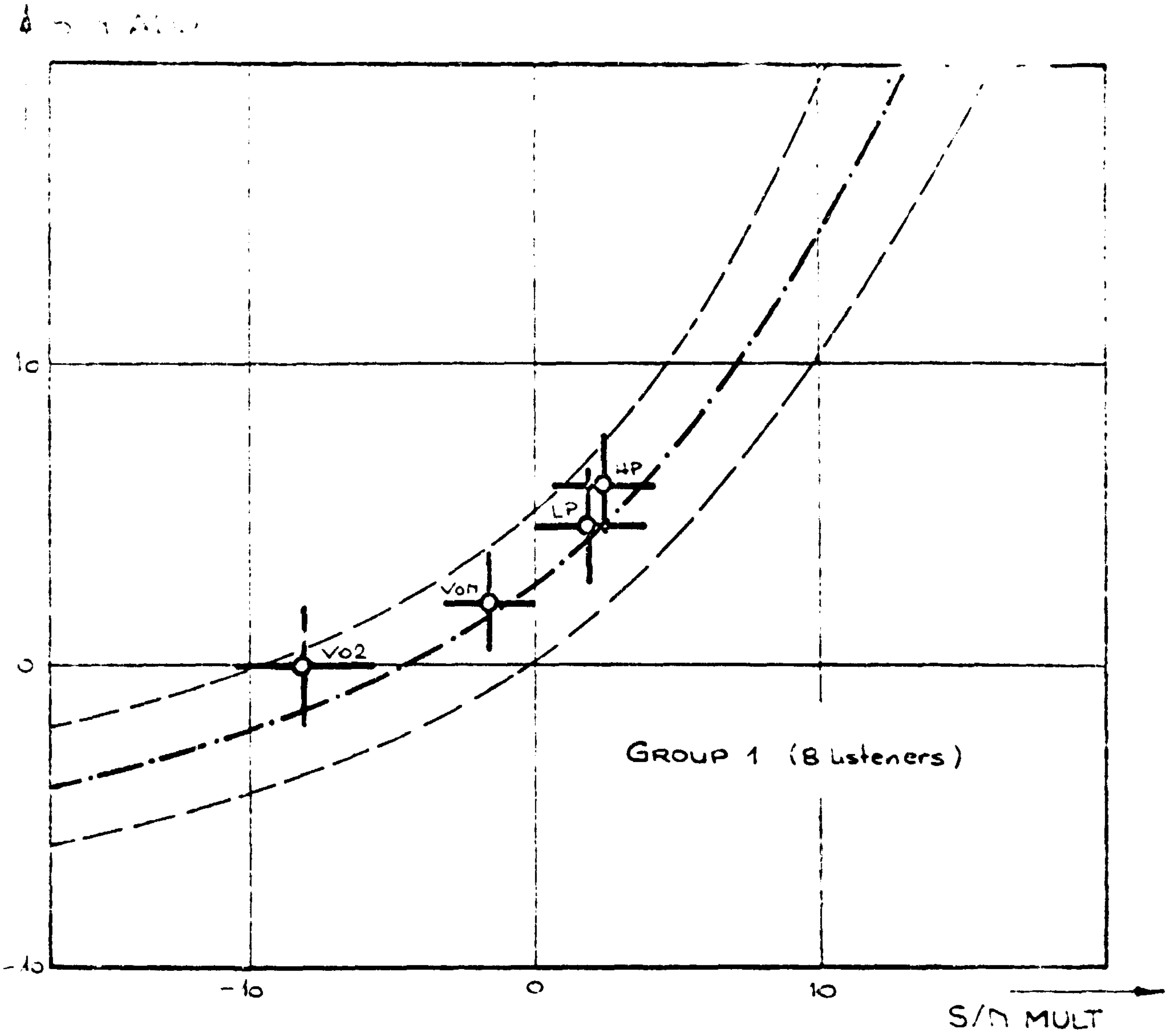


Fig. 3.41

As a comparison to the results of the small trained group on one day, the isopreference points of a large untrained group of about 40 listeners are plotted in Fig. 3.42. The rank order of the test signals is with both reference signals the same as for the trained group and corresponds also with the rank order of the signals which are given by the listeners when asked directly. Both vocoder signals VON and VO2 lie here outside the uncertainty range. One plausible explanation for this effect may be the following. At first the listeners were presented all tests with the multiplicative reference. The unfamiliar character of the vocoder signals caused

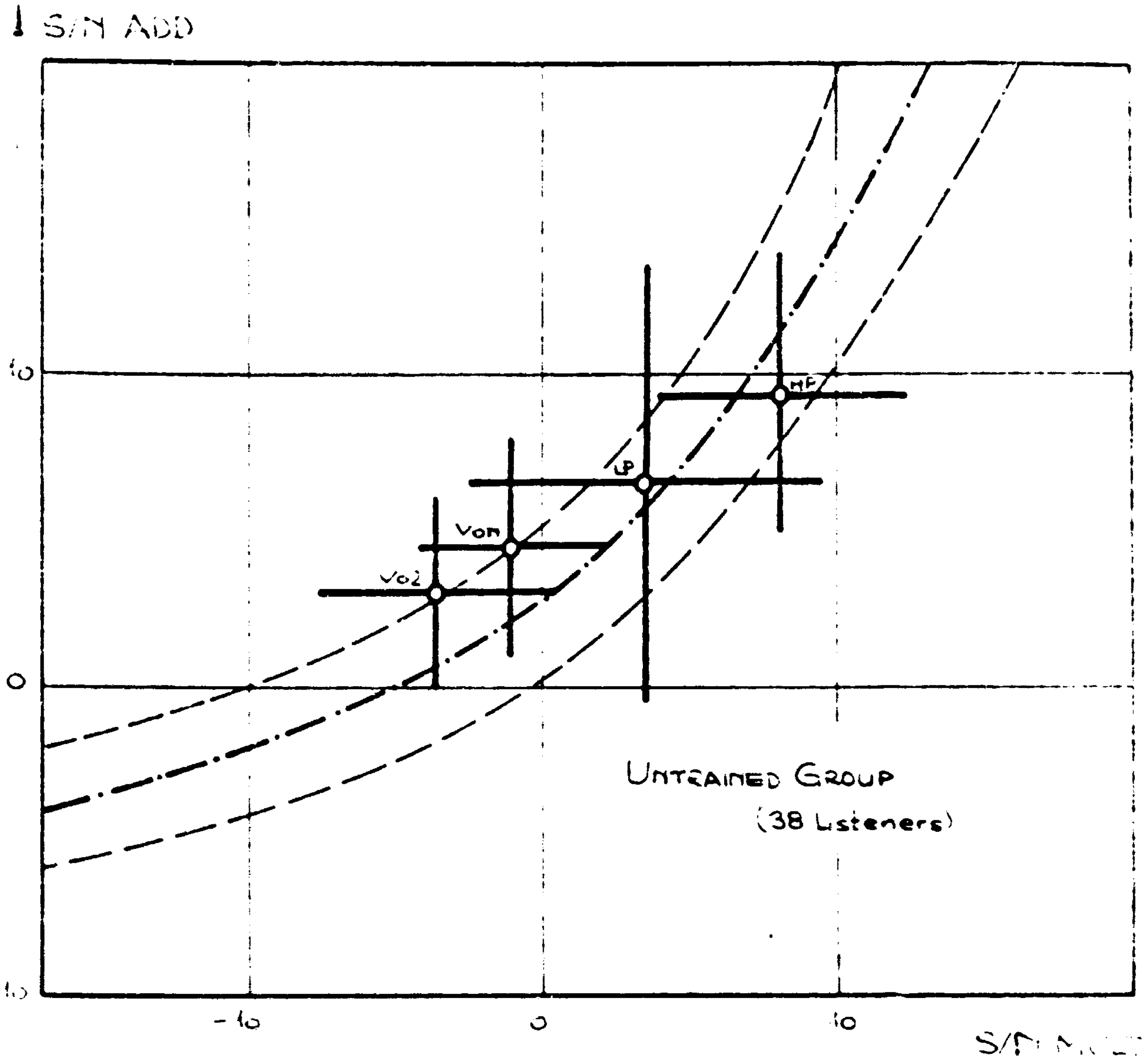


Fig. 3.42

very low isopreference levels. In the subsequent tests with the additive reference the listeners were already somewhat accustomed to the vocoder signals and judged them to be better than before. At the moment we have no other test data to verify the above assumption.

Two further transitivity checks can be made using the isopreference relations determined by area tests in Sec. 3.11 Fig. 3.43 shows the relations between the three reference signals ADD, MULT and DIG. All solid curves are direct measurement results, while the dashed curve ADD-MULT is derived from the two isopreference curves DIG-ADD and DIG-MULT. The good

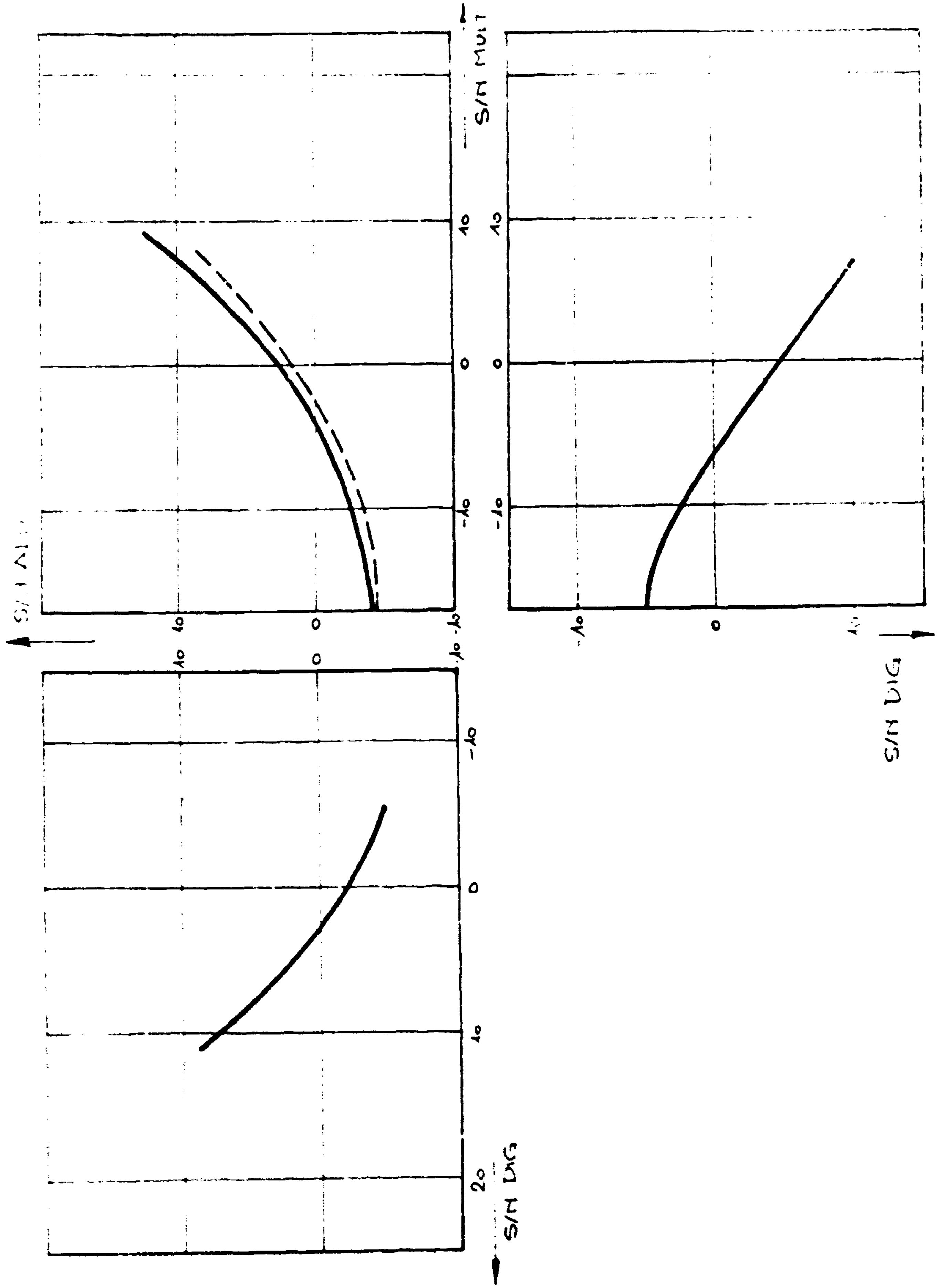


Fig. 3.43

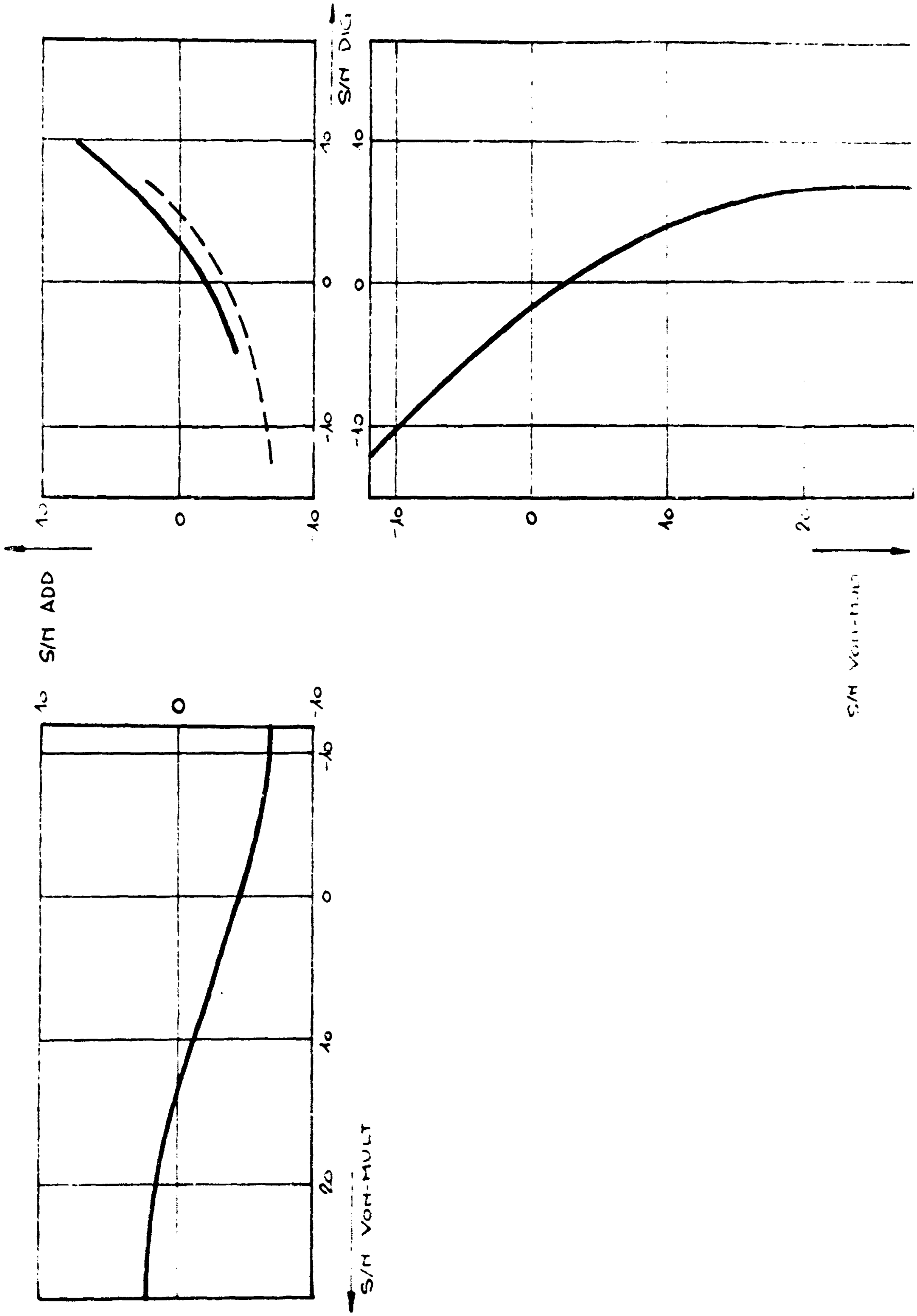

S/N DIG
10
0
-10
S/N ADD
10
0
-10
-10
0
10
20
S/N VON-MULT
S/N VON-MULT
10
0
-10
-10
0
10
20

Fig. 3.44

correspondence between the now two ADD-MULT curves determined by different tests votes also for the transitivity of the reference signals.

In the same way Fig. 3.44 shows the isopreference relations between the two reference signals DIG and ADD and the artificial test signal VON-MULT. The solid curves again are determined by direct tests while the dashed curve ADD-DIG is deduced from the isopreference curves (VON-MULT) - DIG and (VON-MULT) - ADD. The correspondence between the two curves ADD-DIG is not as good as in the example given above. But considering the still unsufficient data which we have collected for the digital reference signal, the dashed curve seems to fit with a reasonable tolerance.

It has been already pointed out that a great part of our preference measurements have been concentrated on test repetitions with the four standard test signals HP, LP, VON and VO2 and the reference signals ADD and MULT. The results discussed in a separate part of Sec. 3.14 cover a period of about two months and show for a single listener and for two different groups the dependence on time of the obtained test results. The examples are taken from five different groups of trained listeners each consisting of about 8 listeners. Now disregarding the dependence on "time" the mean test responses of the single groups shall be calculated. These tests cover a period of about 7 months during which time each group has made 4 - 25 preference tests with all test signals combined with both reference signals. The averaged isopreference points for each of the 5 groups and the four test signals are shown in an ADD versus MULT diagram in Fig. 3.45. Additionally the standard isopreference curve with its uncertainty range can be found in the diagram. The points are widely spread, but lie with only three exceptions within the given uncertainty range.

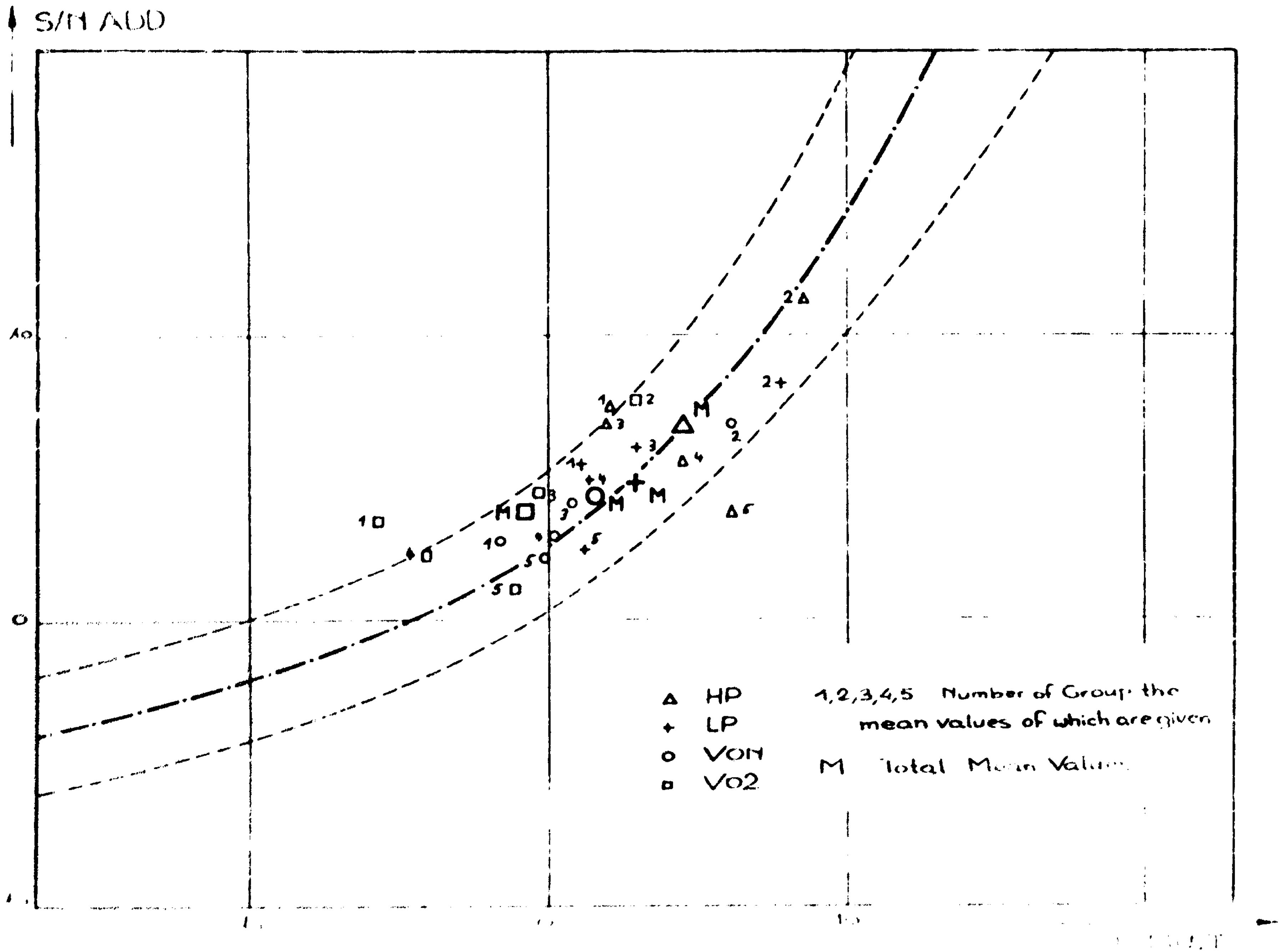

S/H ADD
40
0
1,2,3,4,5 Number of Group the mean values of which are given
M Total Mean Value
HP
LP
VOH
VO2

Fig. 3.45

Training

The available material shows that a human observer has a subjective yardstick of speech quality which changes "randomly" in time and which may be seriously affected by training. We found only very few references in the bibliography with respect to the dependence on time of psycho-acoustic measurements in general, although this effect might have an influence on practically all subjective notions. In standards for intelligibility testing the necessary time for training is defined by the period after which the test scores have reached stationary values. But based upon our preference test data one can see that it is nearly impossible to speak at any time of a "steady state" of the test results. Therefore here we cannot specify a time after which an untrained listener may be qualified as being trained. Further tests will be required to clarify this problem.

Intelligibility and quality

Until recently intelligibility was the only aspect of speech quality which has been used as a criterion for rating speech communication systems in view of their performance. It is a necessary part for the characterization of speech signals but does not give a sufficient description. This becomes obvious when signals with intelligibility scores close to 100 % are compared. We try to describe the speech signal in terms of intelligibility and the relative quality measure "preference".

Intelligibility and preference are by no means independent from each other, for the first term seems to be involved in whole or in part in the second one. A key question of our studies turned out to be the relation between intelligibility and preference. Based upon a body of yet unsufficient data we are not able to answer the question to its full extent. Furthermore it must be mentioned that we studied the intelligibility of isolated words, but used a continuous text for preference testing. Some available results are shown in Fig. 3.46. The diagram shows besides the standard isopreference curve the relation between both reference signals with regard to equal intelligibility. This "isointelligibility" relation has been derived from the measurements in Sec. 3.12. Additionally the results of preference and intelligibility measurements for the test signal HP are given.

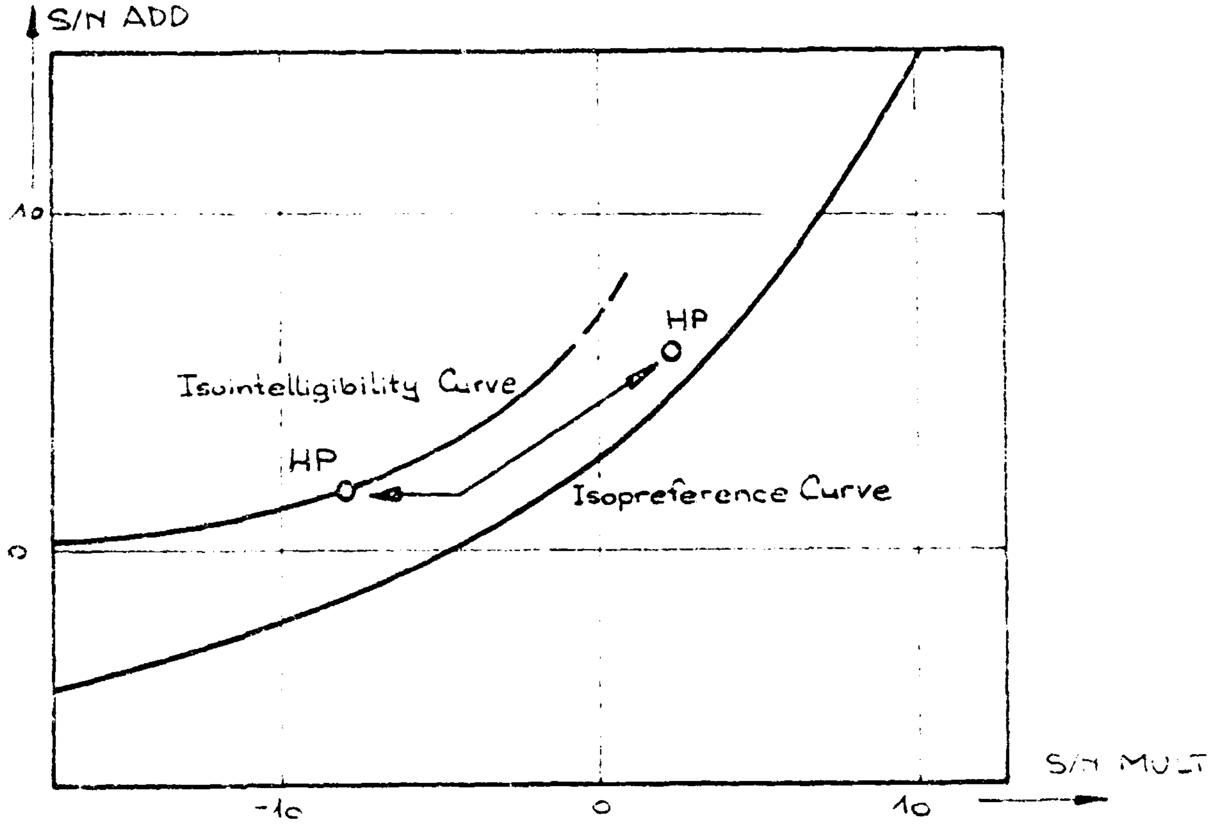


Fig. 3.46

Caused by the very high intelligibility of the multiplicative reference signal, the points of isopreference and isointelligibility of the HP signal are the only corresponding points now available between the two curves. Even from the present insufficient data significant differences between the isopreferent and isointelligible signals may be presumed. Further studies are necessary.

5.5 Conclusions

The following list of brief statements refers to the work of the past and also to the future research period. It tries to summarize where we stand at the moment.

a) Preference tests according to the proposed method are feasible.

b) A close correlation between two types of reference signals could be established namely hifi distorted by adding noise, and anotherone by multiplying the speech signal with noise.

c) Measurements of test signals with the two reference signals give comparable results. Transitivity of preference judgements could be proved to exist with reasonable tolerances.

d) Standardization of a method for preference testing seems to be possible, though we could not yet identify an "ideal" reference signal.

e) The establishing of confidence criteria for single listeners and for groups as function of time and training turned out to be more complex than anticipated.

f) Intelligibility is not simply related to preference. These two criteria may give different rank orders for a set of speech signals.

g) The results of the past period encourage us to continue the subject study along the same general lines.

REFERENCES

/1/ MUNSON, W.A., KARLIN, J.W. : Isopreference Method for Evaluating Speech Transmission Circuits. - J.A.S.A. 34, June 1962, pp. 762 - 774.

/2/ ROTHAUSER, E.H. : Modified Isopreference Method for Audio Quality Measurements. - 66. ASA Meeting, Ann Arbor, Michigan, November 6 - 9, 1963.

/3/ American Standard S 3.2 - 1960, Monosyllabic Word Intelligibility.

/4/ HAHLBROCK, K.H. : Sprachaudiometrie. - G. Thieme Verlag Stuttgart 1957.

/5/ ROTHAUSER, E.H. : A Pulse Excited Vocoder System. - 68. ASA Meeting, Austin, Texas, October 21 - 24, 1964.

/6/ FISZ, M. : Probability Theory and Mathematical Statistics. - John Wiley & Sons, Inc., New York - London.

/7/ FELLER, W. : An Introduction to Probability Theory and its Applications. - John Wiley & Sons, Inc., Now York - London - Sydney.

/8/ HASTINGS jr, C. : Approximations for Digital Computers. - Princeton University Press, Princeton, N.J. 1955.

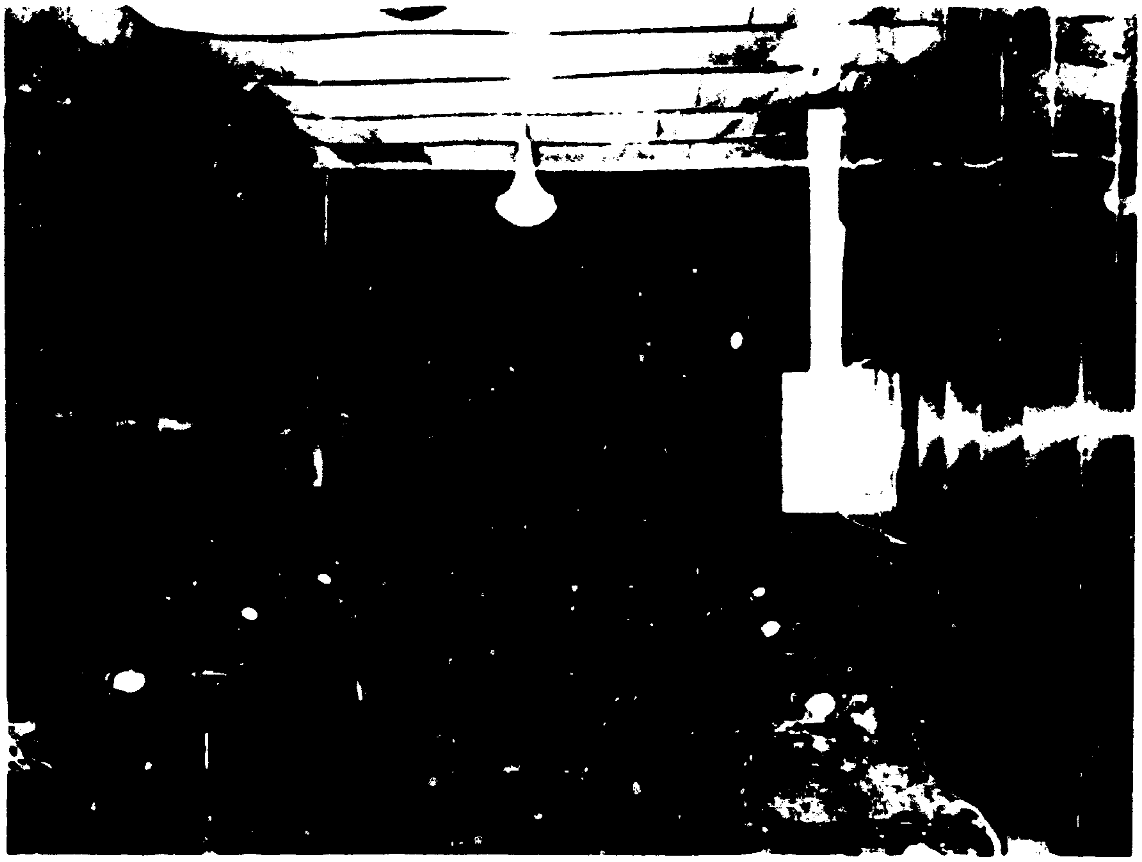
ACKNOWLEDGEMENT

We are indepted to Prof. G. KRAUS, director of the Institut für Niederfrequenztechnik, University of Technology, Vienna for his approval to use the facilities of the institute for the performance of this research and for his continuous encouragement.

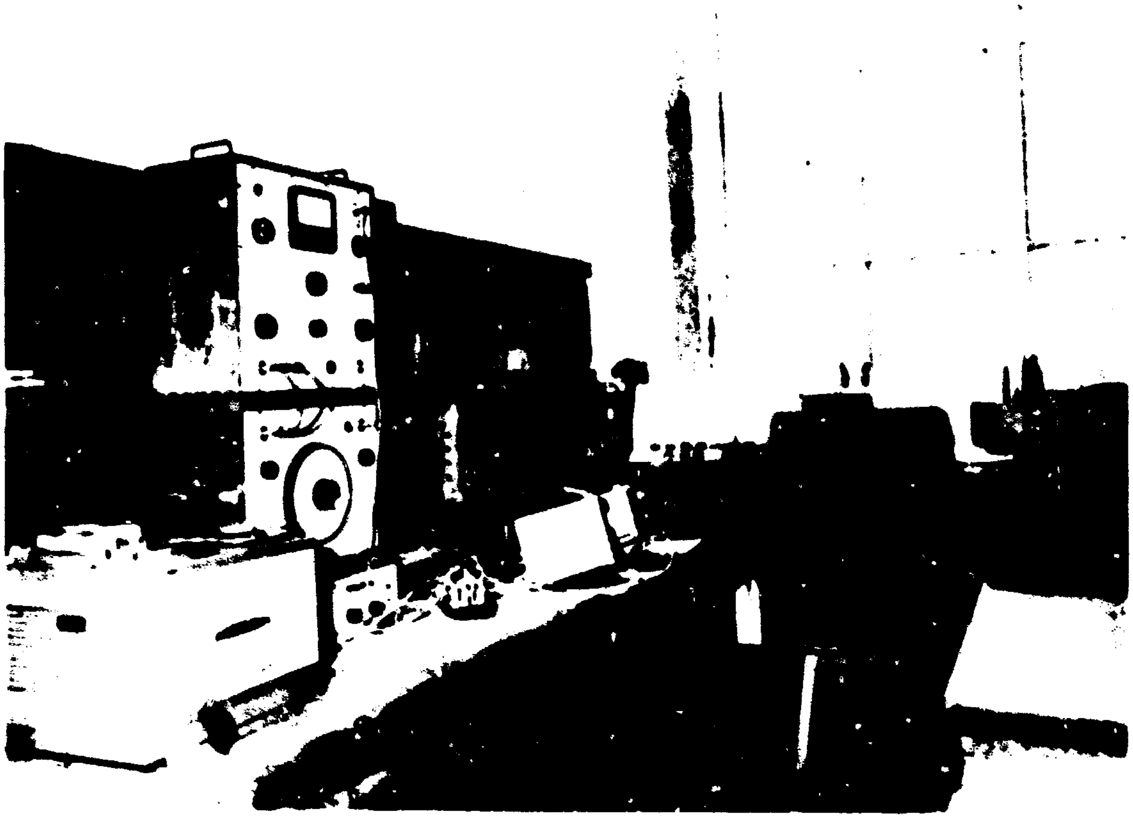
Particularly the authors are grateful to Prof. H. ZEMANEK, manager of the IBM SDD Laboratory Vienna and to Prof. A. P. SPEISER, manager of the IBM Research Laboratory Zürich. Only through their active interest and understanding E.H. ROTHAUSER could participate in the present research work.

APPENDIX

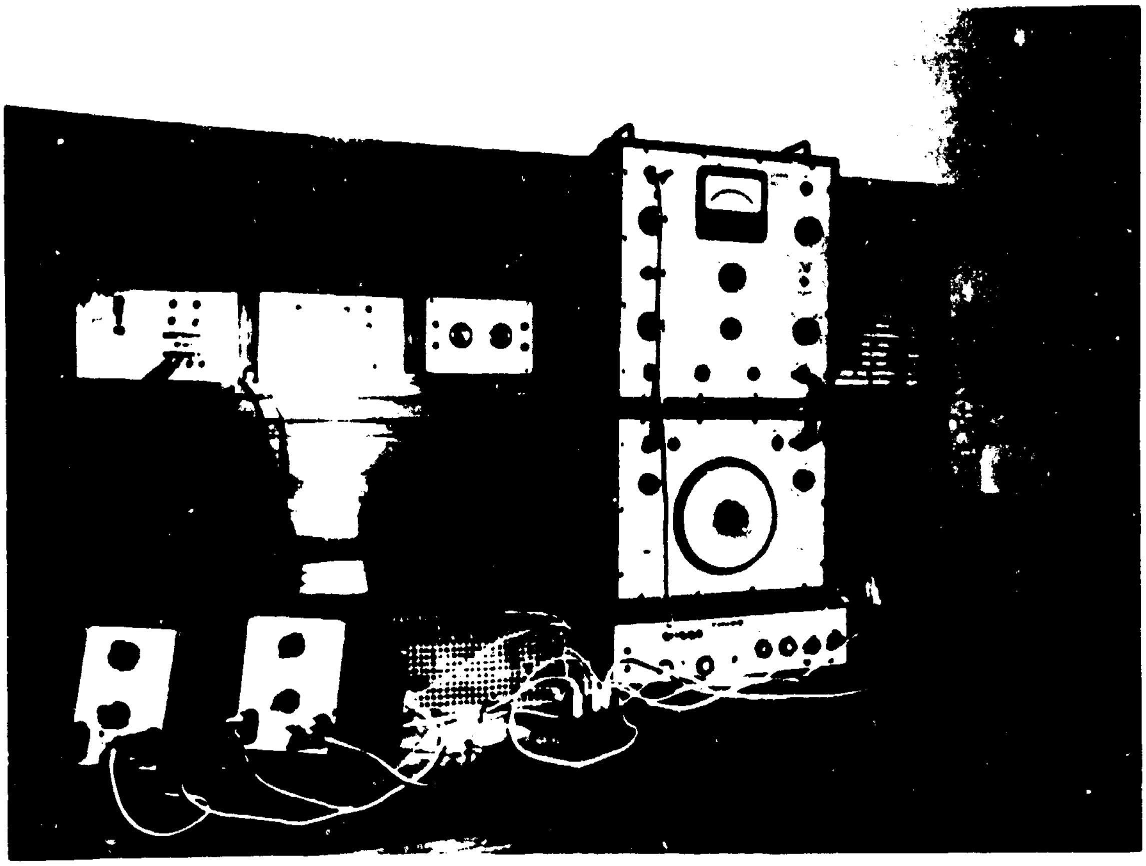
TABLE OF CONTENTS

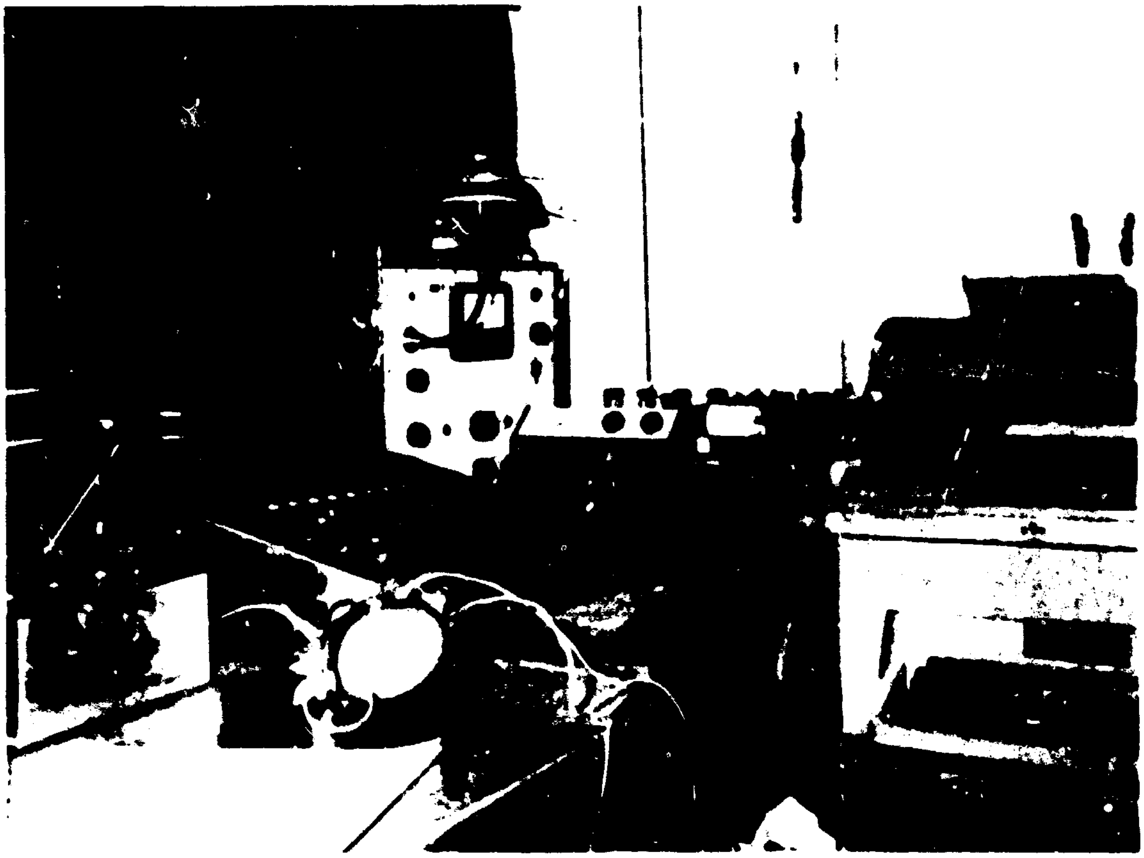


Listener Room



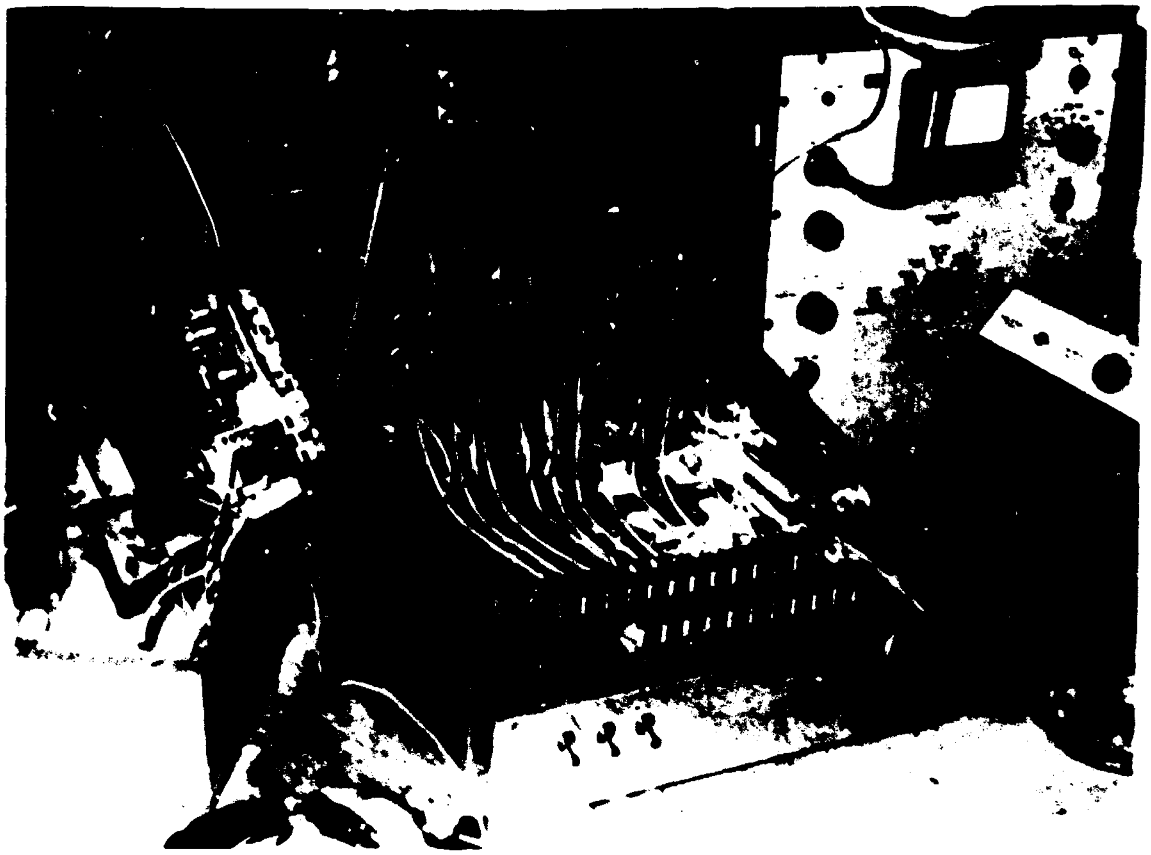
Operator Room



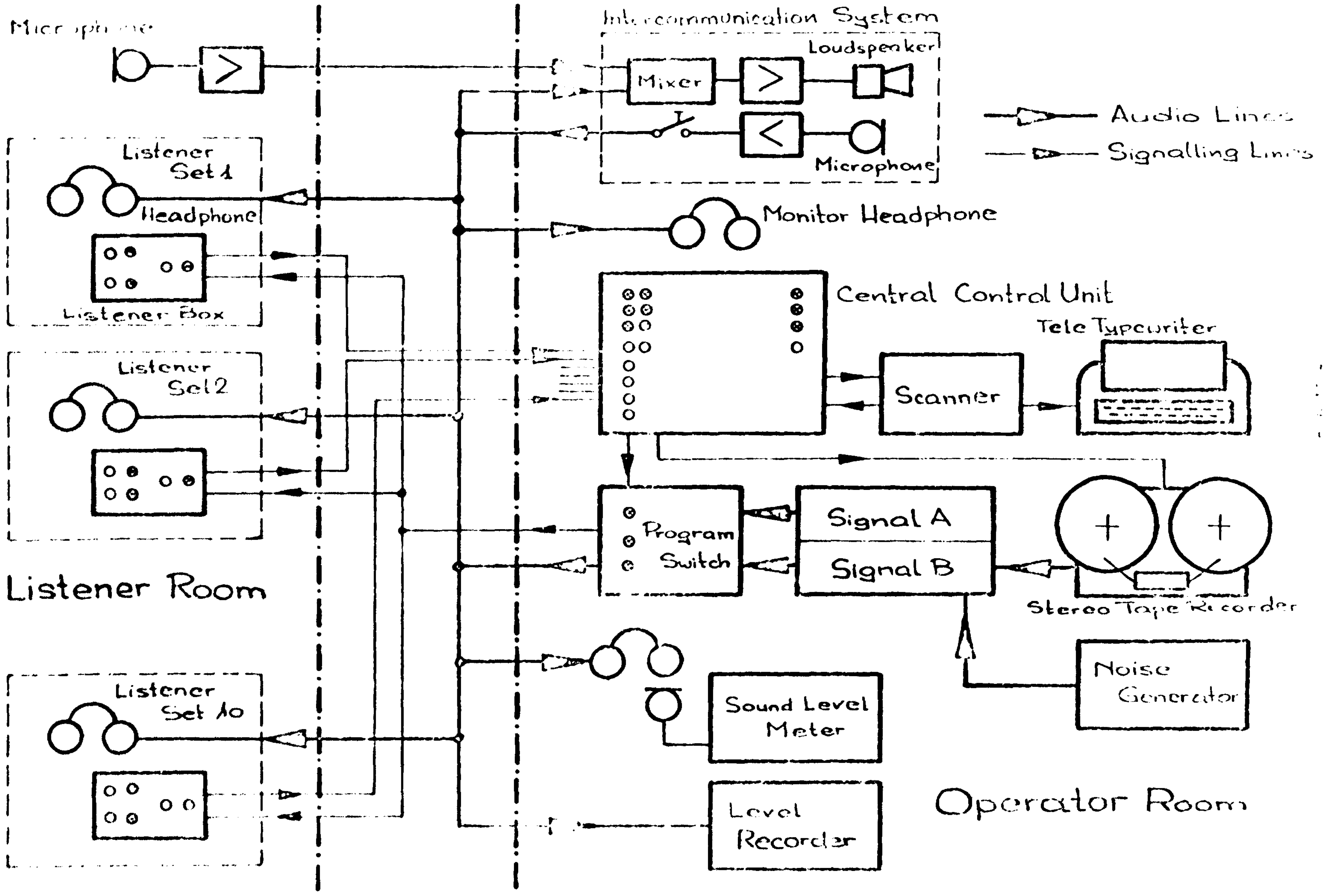


Test Set-up

Listener Set



Central Control Unit


Microphone
Intercommunication System
Mixer
Loudspeaker
Microphone
Audio Lines
Signalling Lines
Listener Set 1
Headphone
Listener Box
Monitor Headphone
Central Control Unit
Tele Typewriter
Scanner
Listener Set 2
Program Switch
Signal A
Signal B
Stereo Tape Recorder
Listener Room
Listener Set 10
Sound Level Meter
Noise Generator
Level Recorder
Operator Room

TEST SET UP

INSTRUMENTATION

REVOX Tape recorder, model G 36
7 1/2 (3 3/4)" per sec. S/N > 55 dB
Frequency response 40 - 18 000 cps +0 -2 dB

Mc INTOSH Power amplifier MC 225
twin-amplifier : 25 watts continuous per channel
18 - 30 000 cps +0/ -0,1 dB

KOSS Professional headphones model No. PRO - 4
30 - 20 000 cps impedance 50 ohms

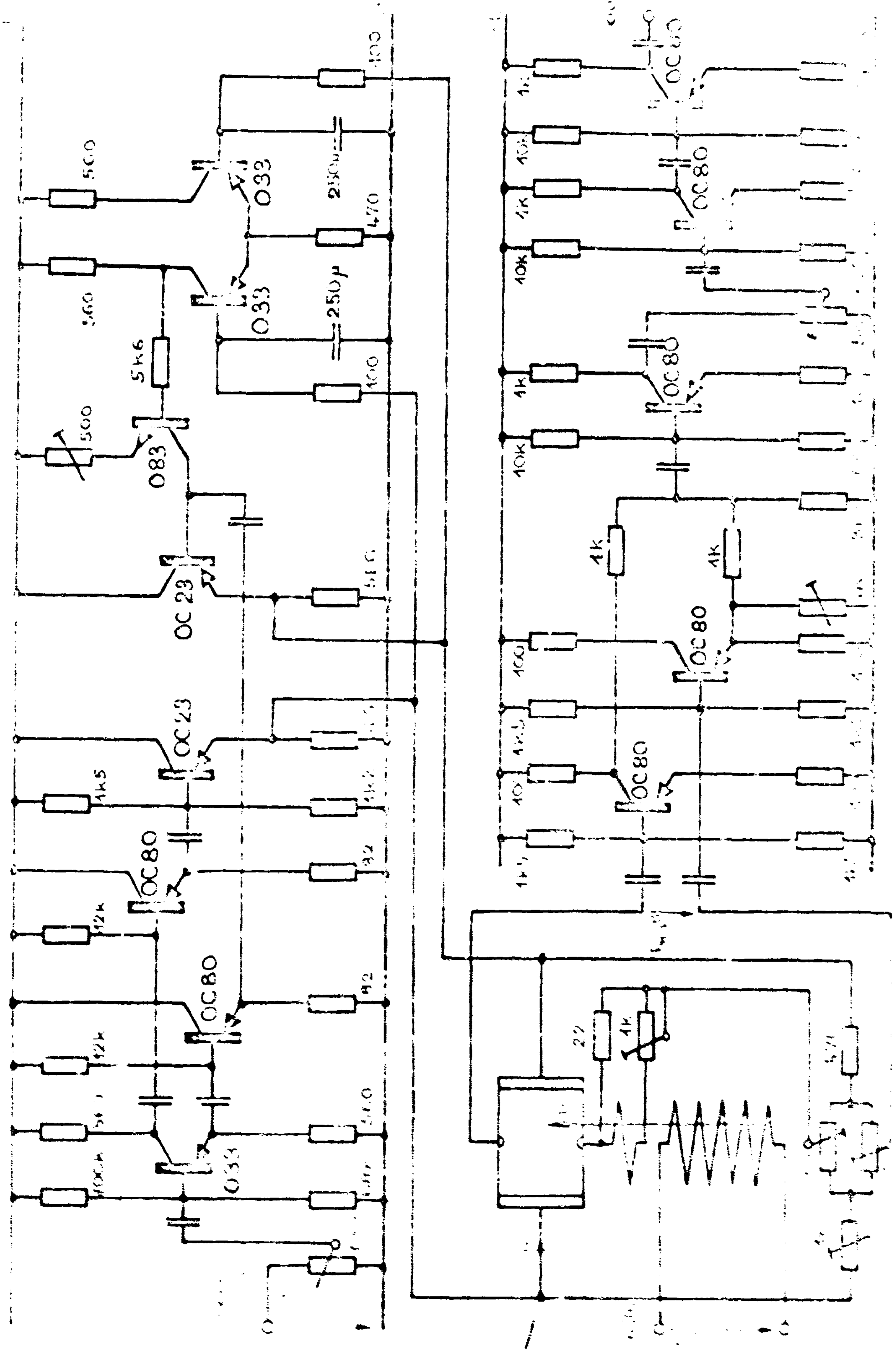
BRÜEL & KJAER Random noise generator type 1402
20 - 20 000 cps linear or - 3 dB/oct weighted
external filter

BRÜEL & KJAER Band pass filter set type 1612
weighting network A, B, C, or Lin and 1/3 oct or 1 oct

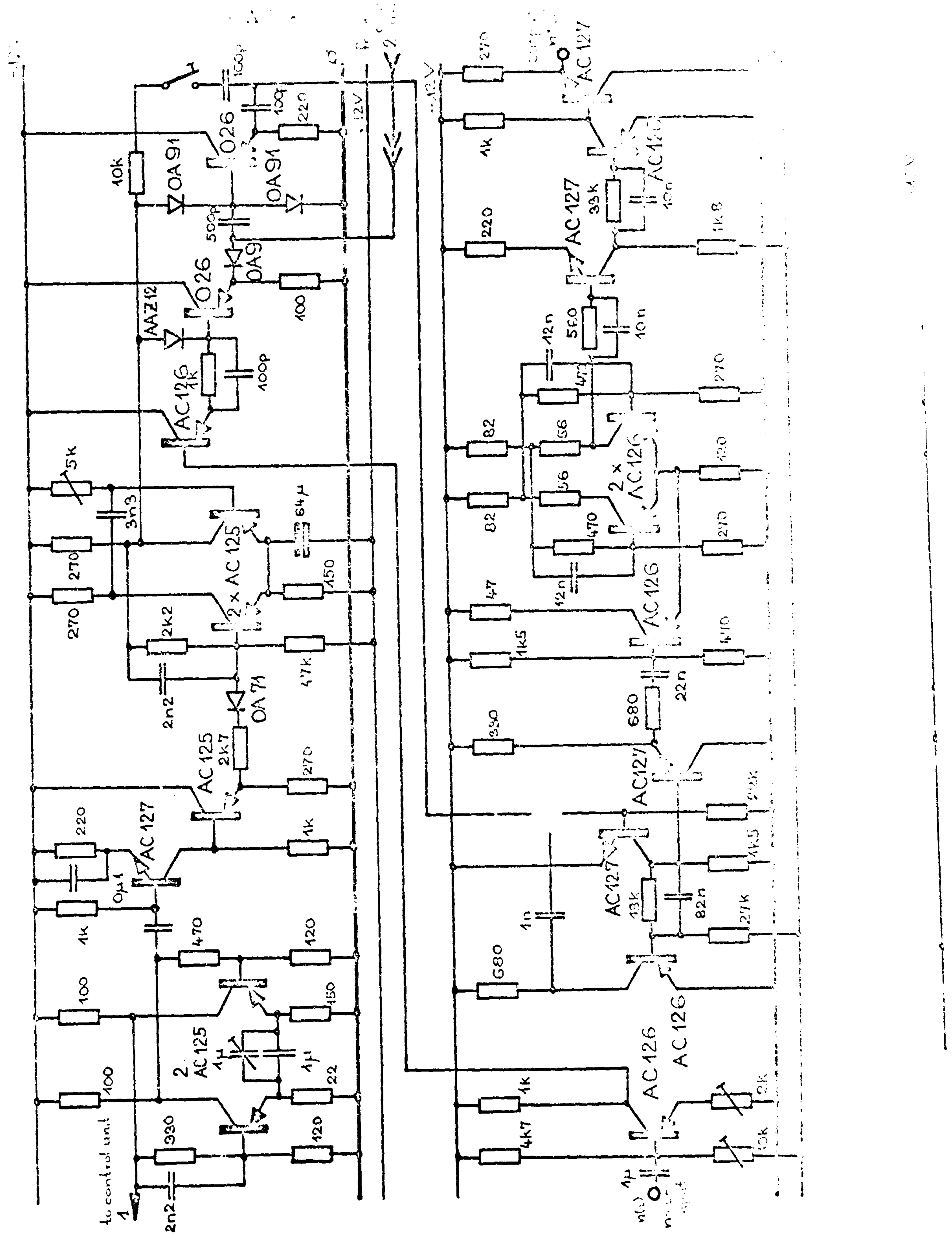
BRÜEL & KJAER Level recorder type 2305
with logarithmic potentiometer 50 dB

BRÜEL & KJAER Microphone amplifier type 2603
Cathode follower type 2614
Condenser microphone type 4[illegible]

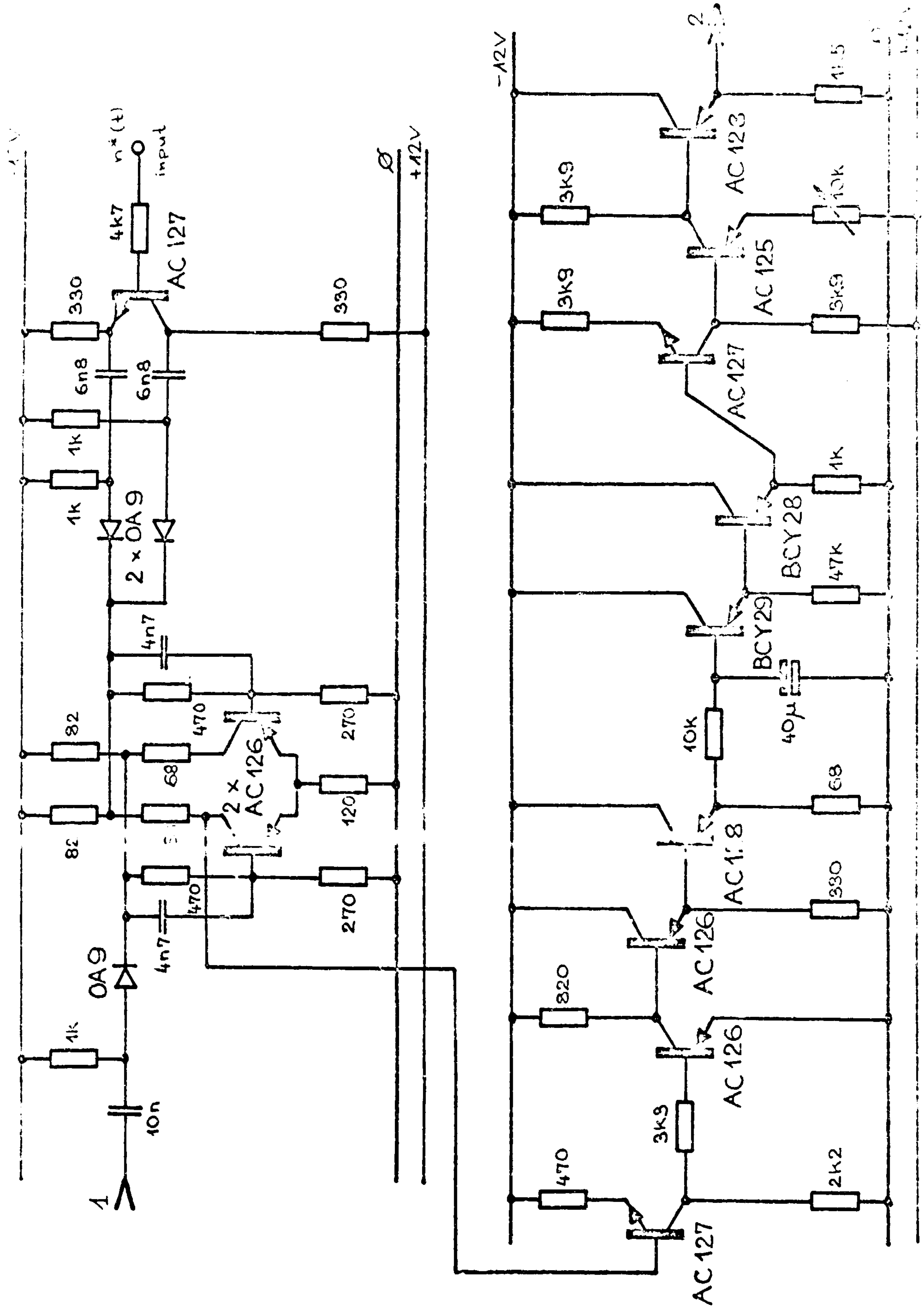
HEWLETT-PACKARD Attenuator set 350 D
0 - 100 dB 600 ohms



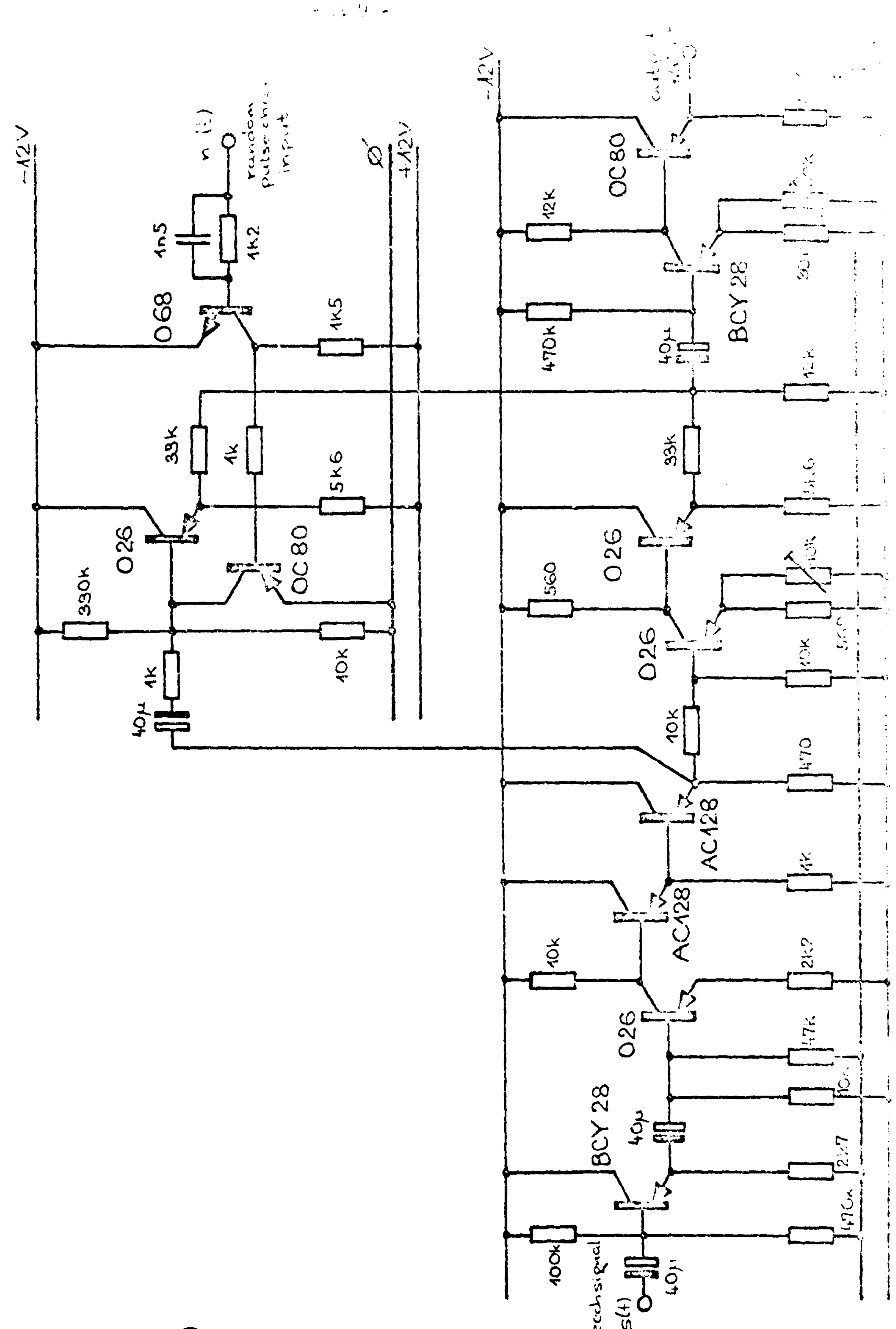
HALL MULTIPLIER



DIG: RANDOM PULSE GENERATOR (continued)



DIG: RANDOM PULSE GENERATOR (CONTROL UNIT)


-12V
n(t)
random pulsechain input
1n5
1k2
O68
1k5
Ø
+12V
33k
1k
5k6
O26
OC80
330k
10k
1k
40µ
-12V
output
OC80
12k
BCY 28
470k
40µ
12k
33k
5k6
O26
560
10k
O26
10k
10k
470
AC128
1k
AC128
10k
2k2
O26
47k
10k
BCY 28
40µ
2k7
100k
470k
40µ
speechsignal s(t)

DiG: SWITCH

PROGRAM SWITCH

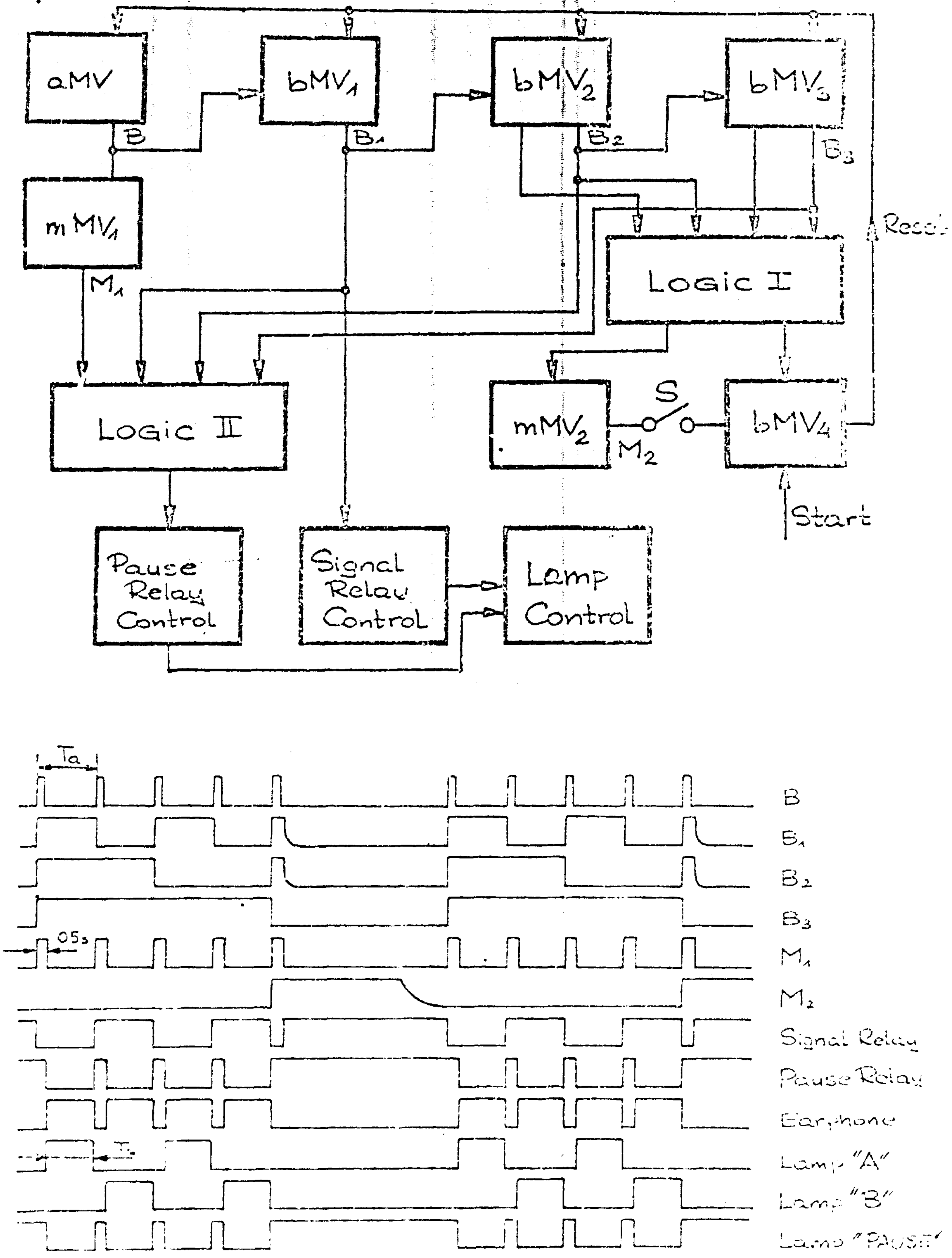
The program switch serves to get the desired mode of signal presentation. Block diagram and corresponding time plans are shown on the next page. The duration of presentation T_s is equal for both speech signals A and B and is given by the formula :

$$T_s = T_a - 0,5 \text{ (sec)}$$

where T_a is the pericd of the astable multivibrator aMV. The earphones are cut off for 0,5 seconds whenever speech signals are switched over. This short pause corresponds to the period of the monostable multivibrator mMV_1 which causes the pick-up of the pause relay over circuit "logic II". The switch-over of the speech signals is done by the signal relay which is controlled by the bistable multivibrator bMV_1. The counting circuit bMV_2 and bMV_3 counts the presented signal pairs. Eligibly after one, two or three presented signal pairs the circuit "logic I" is able to excite bMV_4 which produces the reset of aMV_1, bMV_1, bMV_2, and bMV_3 and over logic II the cut-off of the earphones by the pause relay. After a starter impulse the flipping bMV_4 opens aMV to bMV_3 and the process is repeated. Switching of S enables an automatic control of the pause duration by mMV_2. Three small lamps "A", "B", and "PAUSE" are controlled by the signal relay and the pause relay and indicate the program run optically.

The following modes of presentation are possible :

1. One, two or three successive presentations of the signal pair A-B
2. The duration T_s of presenting the speech signals A and B is equal but can be varied from 2 to 15 seconds.
3. The duration of the long pause between two repeated signal pairs ABAB can be varied from 4 to 20 seconds.



PROGRAM SWITCH

Attenuator Setting	TEST 1									
	DRK	DRU	MUE	SVE	SPA	KLE	BER	PRA	LER	ZLA
6										
7										
8										
9										
10										
11	○							○	○	
12								○	○	
13									○	
14			○						○	○
15								○	○	
16		○	○	○			○	○	○	○
17	○	○	○	○				○		○
18		○	○	○			○			
19	○		○	○						
20	○		○	○	○	○	○			
21	○	○		○			○			○
22	○	○	○	○	○	○				○
23	○	○	○		○	○				○
24	○	○	○	○	○	○				○
25	○	○	○	○	○	○		○	○	

Attenuator Setting	TEST 2									
	DRK	DRU	MUE	SVE	SPA	KLE	BER	PRA	LER	ZLA
6										
7										
8										
9										
10										
11										
12									○	
13							○		○	
14							○		○	
15				○					○	○
16	○	○	○				○	○	○	○
17	○			○	○		○	○	○	
18		○	○			○	○	○	○	
19		○	○	○	○		○	○	○	
20	○	○		○	○				○	○
21		○	○	○	○	○	○	○	○	○
22	○	○	○	○	○	○	○			
23		○	○	○	○	○	○	○		○
24	○	○	○	○	○	○	○	○	○	
25	○	○	○	○	○	○	○	○	○	○

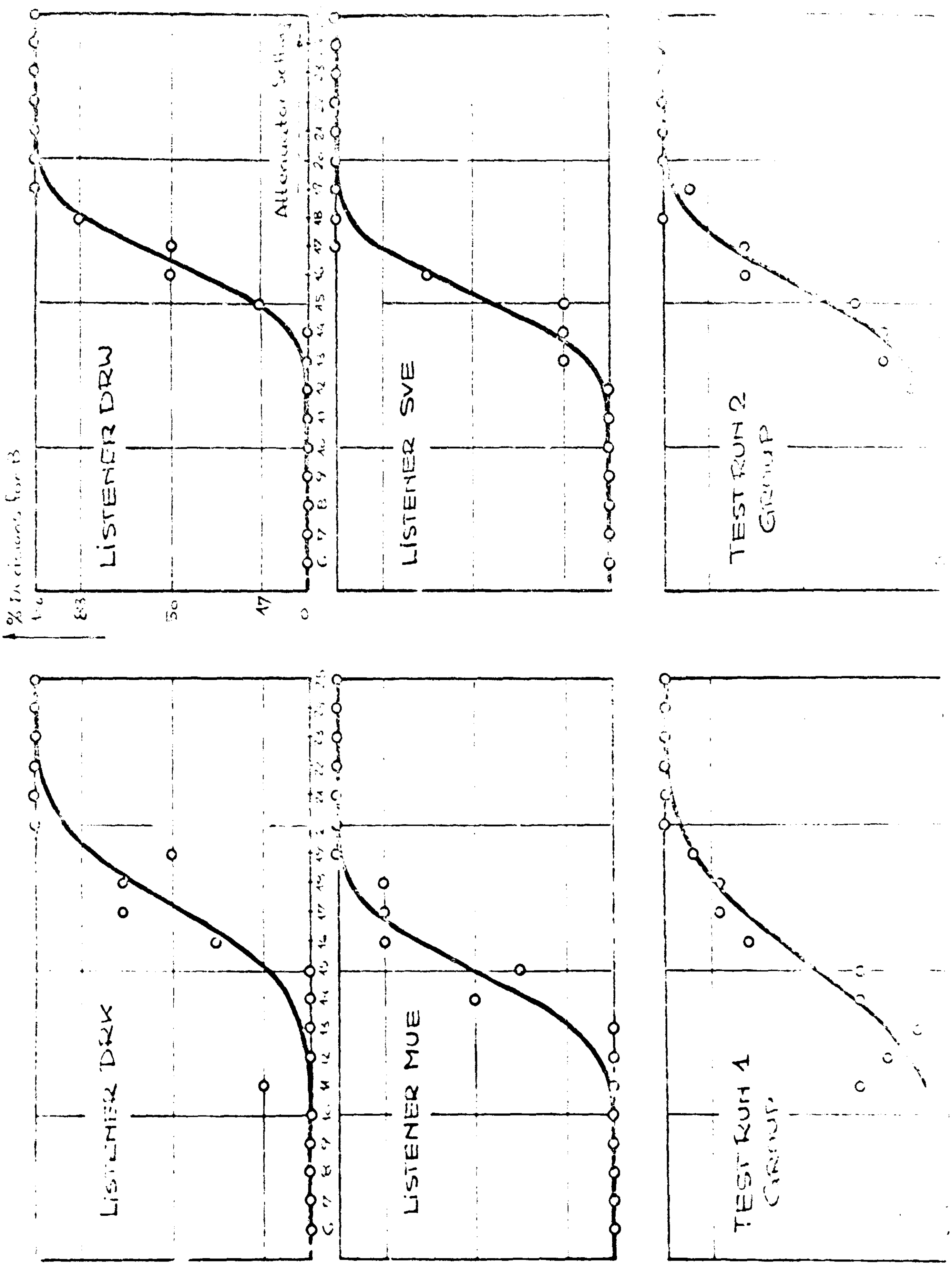
Attenuator Setting	TEST 3									
	DRK	DRU	MUE	SVE	SPA	KLE	BER	PRA	LER	ZLA
6										
7										
8										
9										
10										
11										
12										
13			○							
14		○	○							
15		○								
16		○								
17	○	○								
18	○	○	○							
19	○	○	○							
20	○	○	○							
21	○	○	○							
22	○	○	○							
23	○		○							
24	○		○							
25	○	○	○							

Attenuator Setting	TEST 4									
	DRK	DRU	MUE	SVE	SPA	KLE	BER	PRA	LER	ZLA
6										
7										
8										
9										
10										
11										
12										
13										
14										
15			○						○	
16	○		○	○					○	○
17			○	○			○		○	○
18	○			○	○		○	○	○	○
19		○	○	○	○		○	○		
20	○	○	○	○	○	○	○			
21	○	○	○	○	○	○	○	○	○	○
22	○	○	○	○	○	○	○	○	○	○
23	○	○	○	○	○	○	○	○	○	○
24	○	○	○	○	○	○	○	○	○	○
25	○	○	○	○	○	○	○	○	○	○

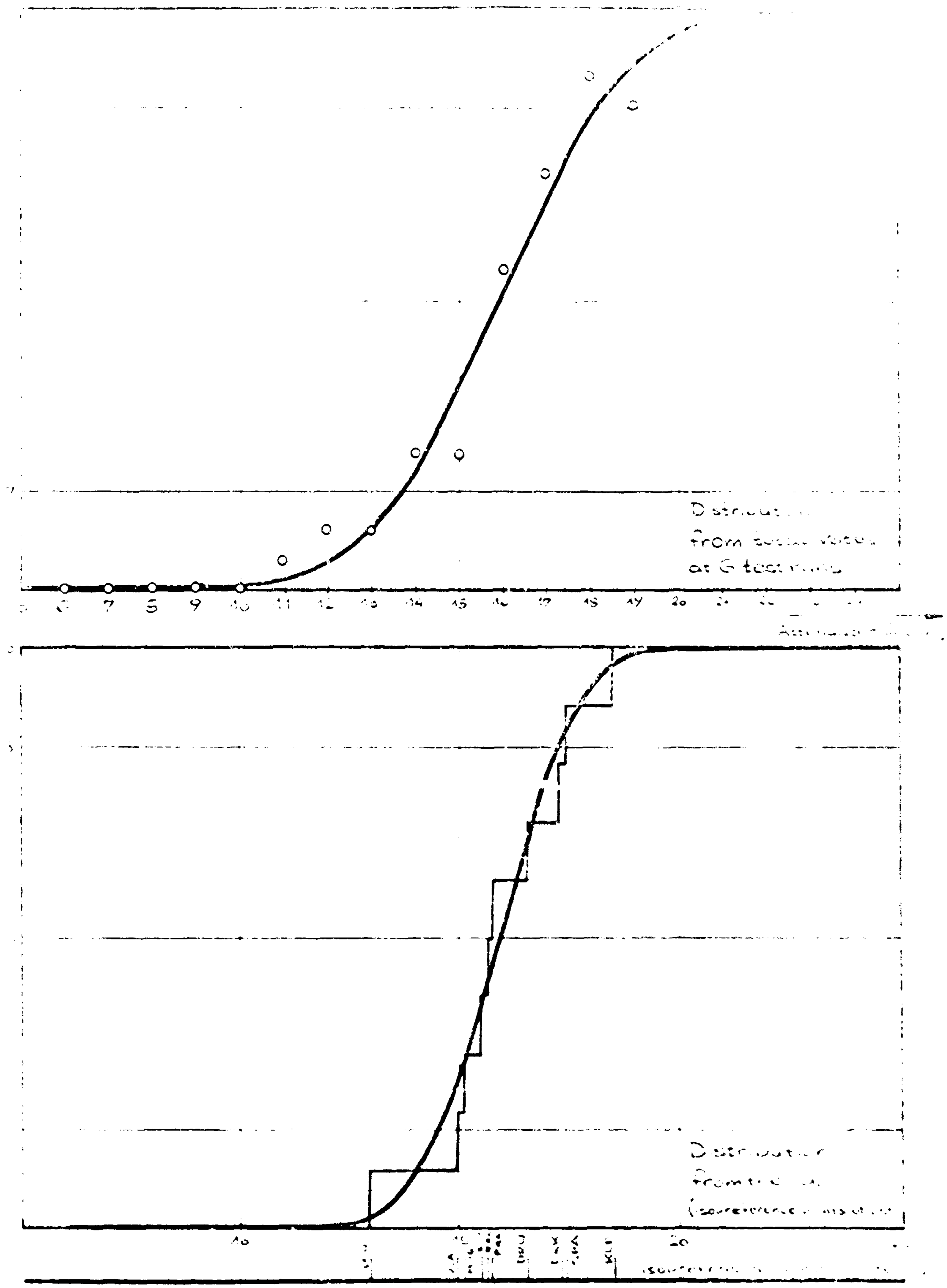
Attenuator Setting	TEST 5									
	DRK	DRU	MUE	SVE	SPA	KLE	BER	PRA	LER	ZLA
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										
16										
17	○		○	○			○			○
18	○	○	○	○	○	○	○		○	○
19		○	○	○	○		○		○	○
20	○	○	○	○	○	○	○	○		○
21	○	○	○	○	○	○	○	○	○	○
22	○	○	○	○	○	○	○	○	○	○
23	○	○	○	○	○	○	○	○	○	○
24	○	○	○	○	○	○	○	○	○	○
25	○	○	○	○	○	○	○	○	○	○

Attenuator Setting	TEST 6									
	DRK	DRU	MUE	SVE	SPA	KLE	BER	PRA	LER	ZLA
6										
7										
8										
9										
10										
11										
12										
13										
14			○							
15										
16			○	○						
17		○	○	○						
18		○	○	○						
19		○		○						
20	○	○	○							
21	○	○	○							
22	○	○	○							
23	○	○	○	○						
24	○	○								
25	○	○	○	○						○

REPEATED PREFERENCE TEST : COMPLETE TEST DATA



REPEATED PREFERENCE TEST VON/MULT
NORMAL DISTRIBUTIONS FOR LISTENERS AND GROUP



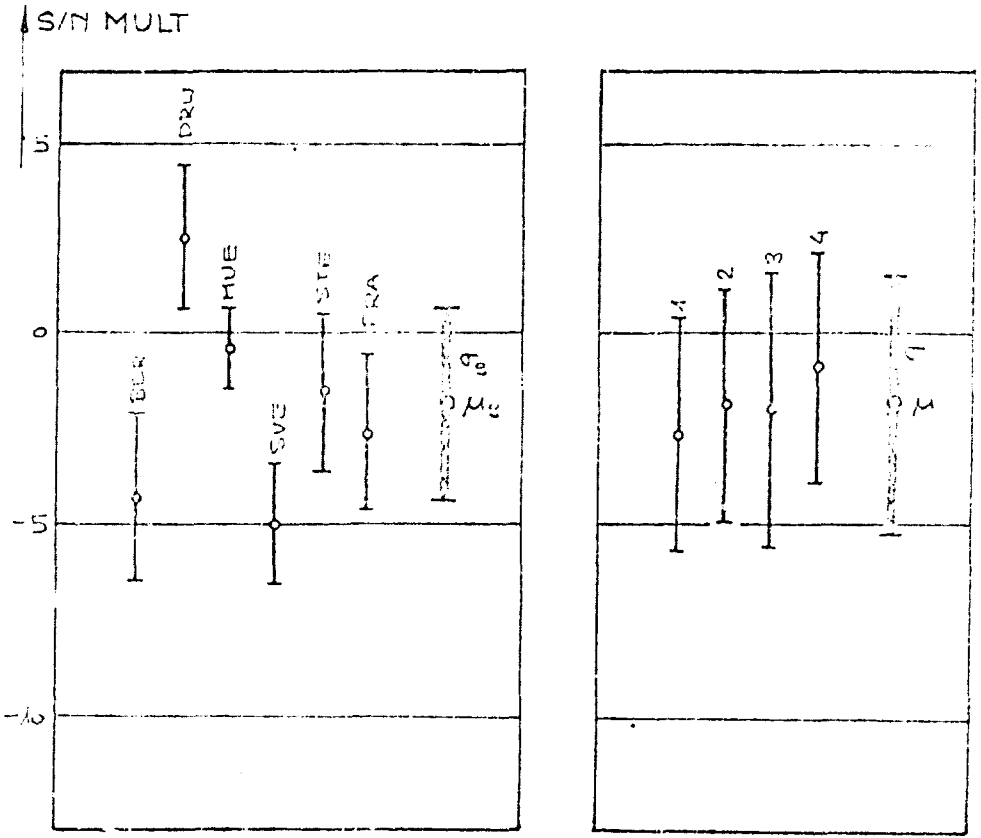
REPEATED PREFERENCE TEST NORMAL DISTRIBUTIONS FOR GROUP

List	μ_i	σ_i
BER	- 4.3	2.2
DRW	2.6	1.9
MUE	- 0.5	1.1
SVE	- 5	1.6
STE	- 1.5	2.1
PRA	- 2.6	2.0

Test	μ	σ
1	-2.7	3.0
2	-1.9	3.1
3	-2.0	3.4
4	-0.9	3.0

13.05.1965 Vom/Mult

$\mu_g = -1.9$ dB $\mu = -1.8$ dB
$\sigma_g = 2.5$ dB $\sigma = 3.4$ dB



Repeated Preference Test Vom/Mult

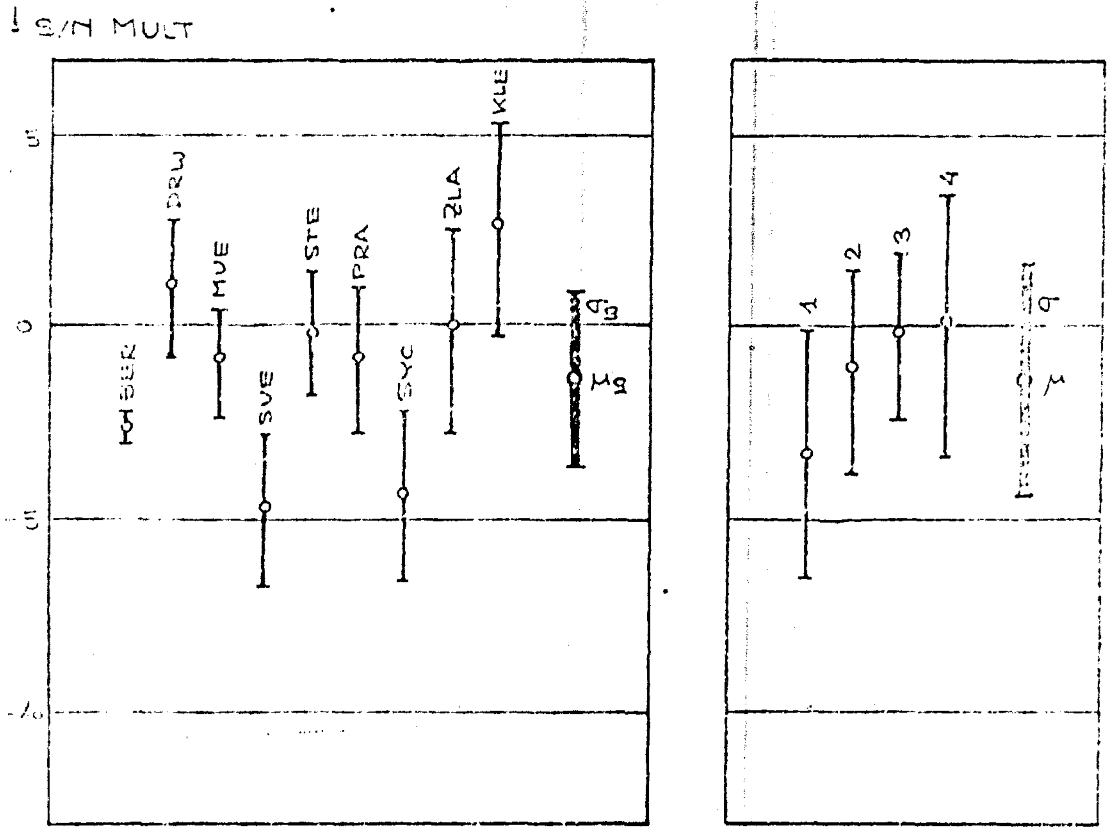
List	μ_i	σ_i
BER	-2.6	0.4
DRW	1.1	1.9
MUE	-0.9	1.7
SVE	-4.7	2.0
STE	-0.2	1.6
PRA	-0.8	2.2
SYC	-4.3	2.6
ZLA	0	2.7
KLE	2.7	2.8

TEST NR	μ	σ
1	-3.3	3.2
2	-1.1	2.6
3	-0.2	2.4
4	0.1	3.4

14.05.1965 VON/MULT

$\mu_g = -1.1$ dB $\quad \mu = -1.4$ dB

$\sigma_g = 2.1$ dB $\quad \sigma = 3.0$ dB



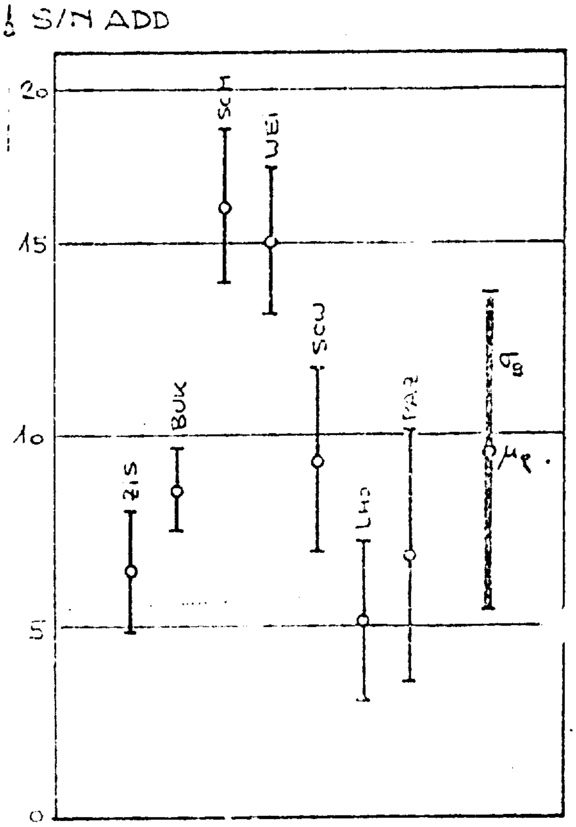
REPEATED PREFERENCE TEST VON / MULT

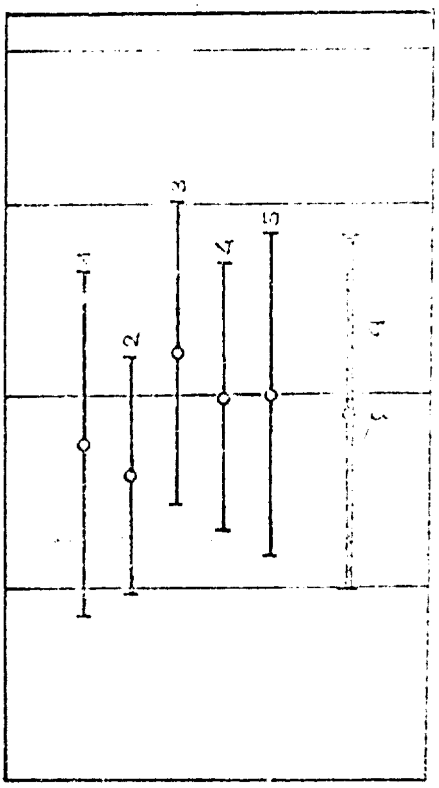
List	μ_i	σ_i
ZIS	6.4	1.6
BUK	8.5	1.1
SCM	15.9	2.0
WEI	15.0	1.9
SCW	9.3	2.4
LHO	5.1	2.1
PAZ	6.8	3.3

Test	μ	σ
1	8.7	[illegible]
2	7.9	3.1
3	11.1	[illegible]
4	9.9	[illegible]
5	10.0	4.2

02.06.1965 TP/ADD

$\mu_g = 9.5\,dB$ $\mu = 9.5\,dB$
$\sigma_g = 4.2\,dB$ $\sigma = 4.6\,dB$





REPEATED PREFERENCE TEST TP / ADD

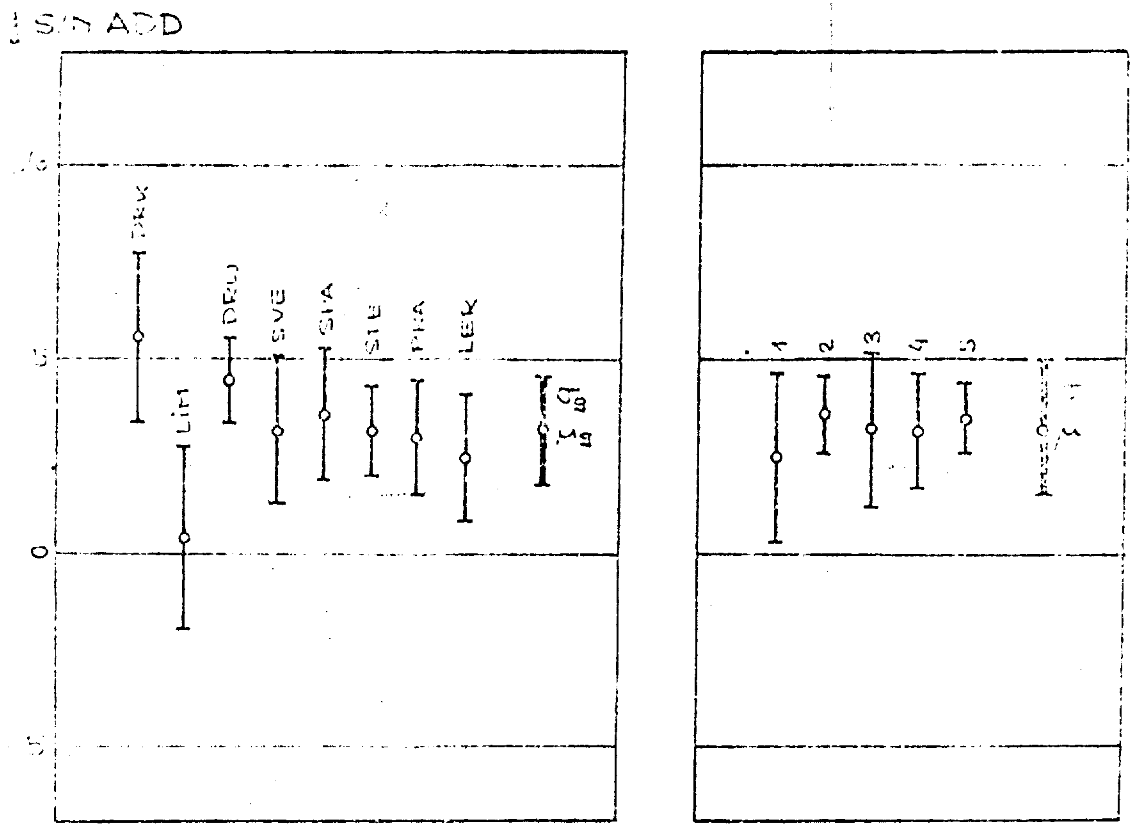
List	μ	σ
DRK	5.6	2.2
LIM	0.4	2.4
DRU	4.5	1.1
SVE	3.2	1.9
STA	3.7	1.7
STE	3.2	1.2
PRA	3.0	1.5
LEK	2.5	1.7

Test	μ	
1	2.5	[illegible]
2	3.7	[illegible]
3	3.3	[illegible]
4	3.2	1.5
5	3.5	0.9

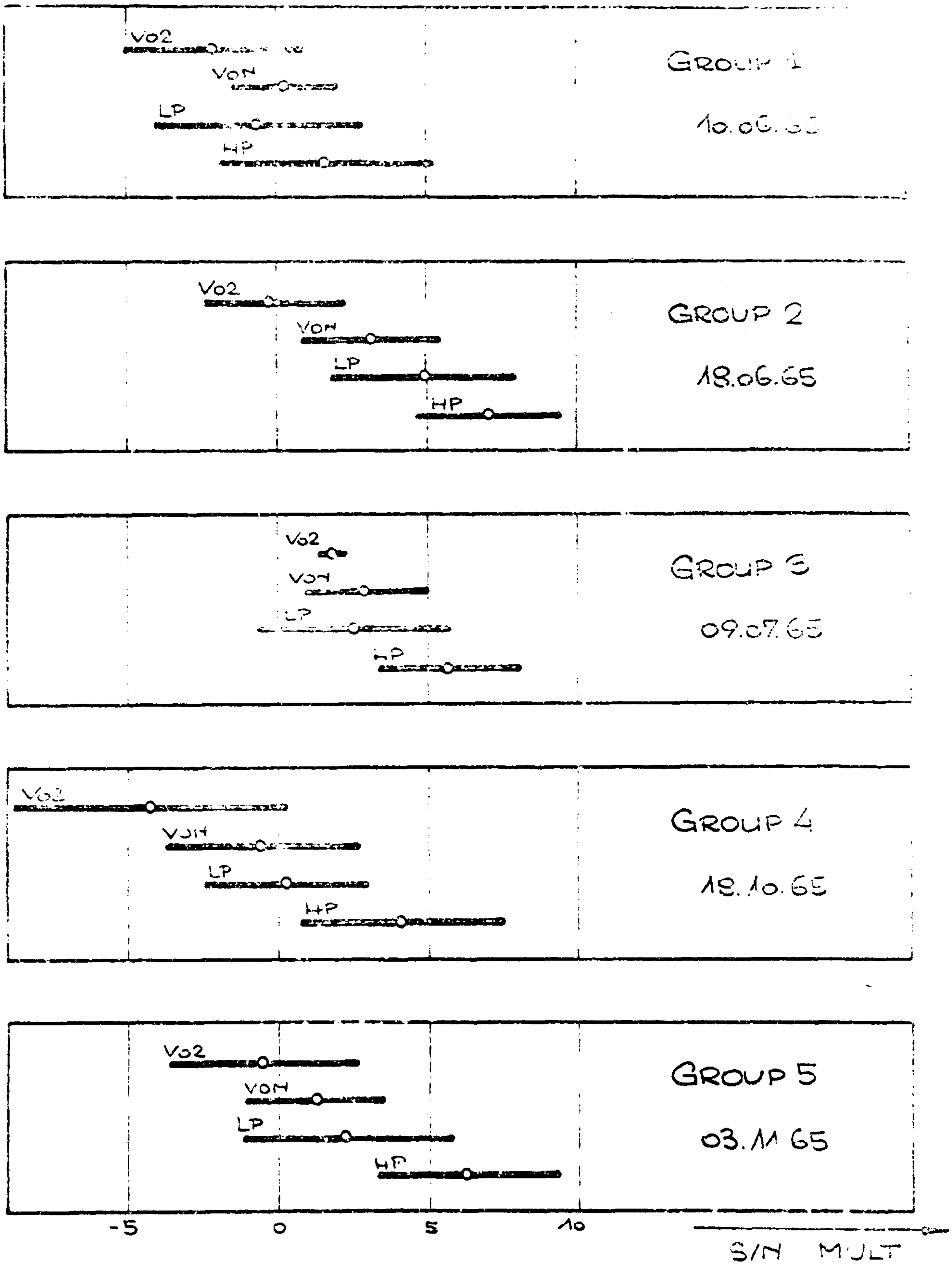
03. 06. 1965 VON /ADD

$\mu_g = 3.3$ dB $\mu = 3.2$ dB

$\sigma_g = 1.4$ dB $\sigma = 1.7$ dB

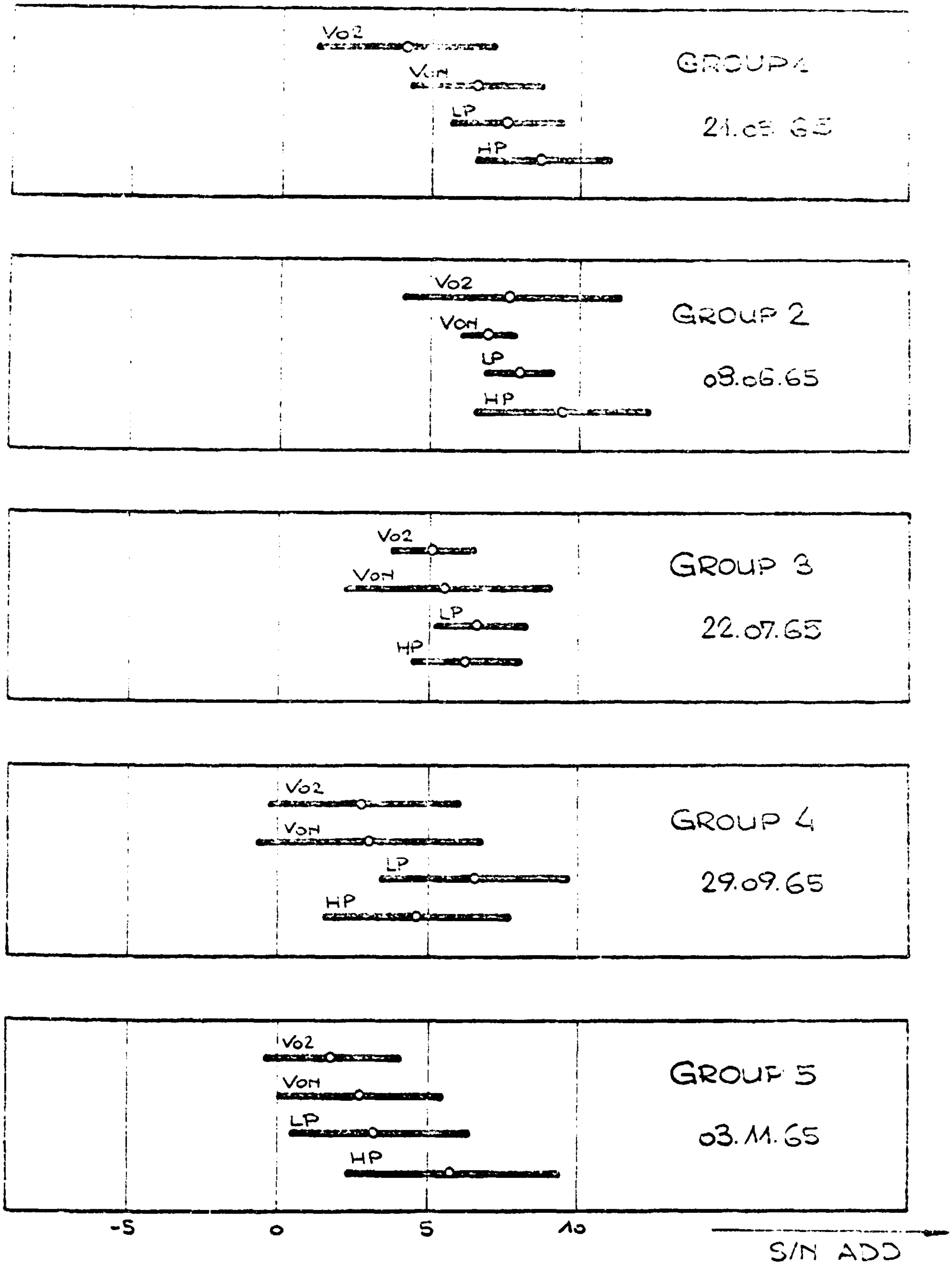


REPEATED PREFERENCE TEST VON / ADD



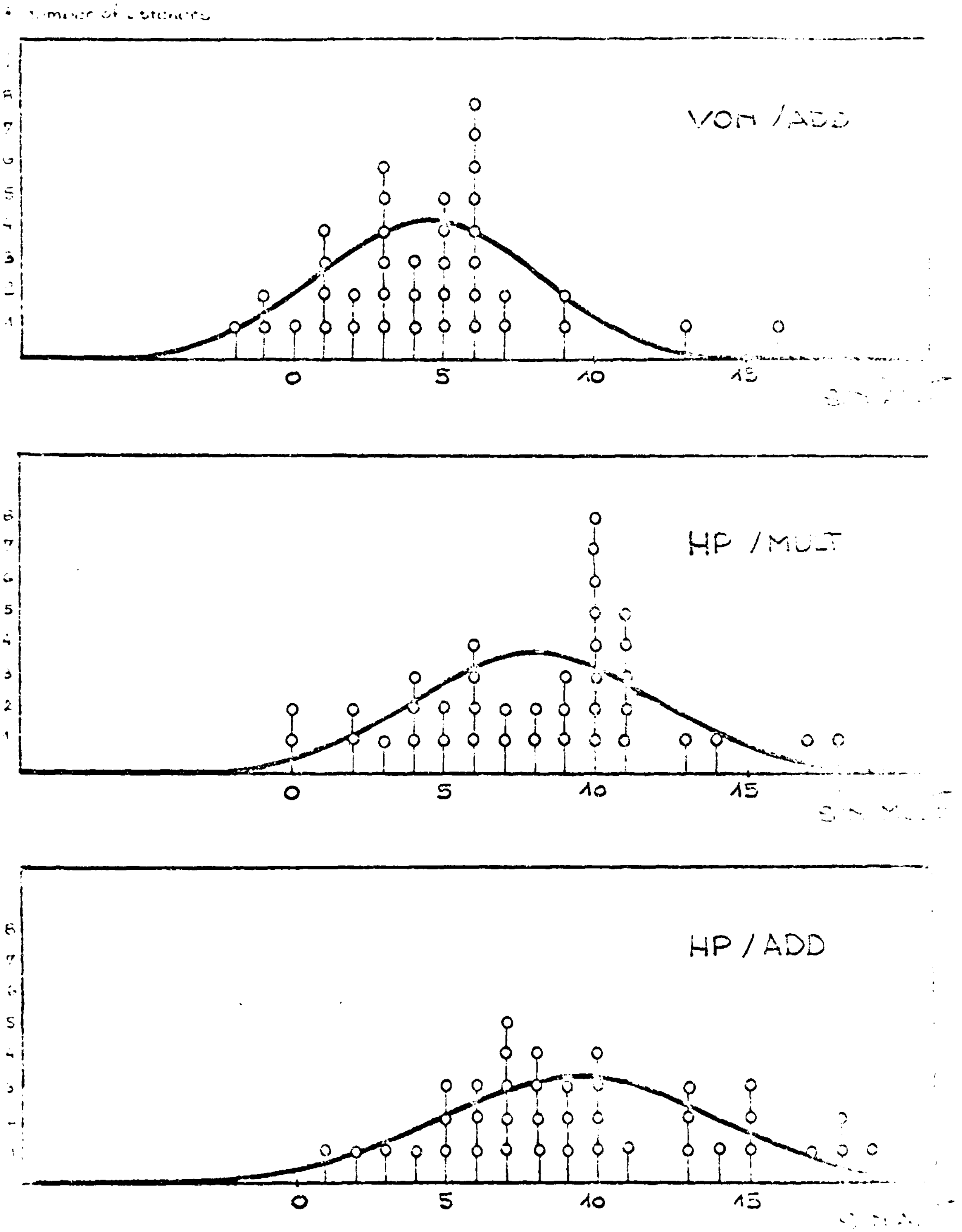
RANK-ORDERING BY PREFERENCE TESTS

MULTIPLICATIVE REFERENCE

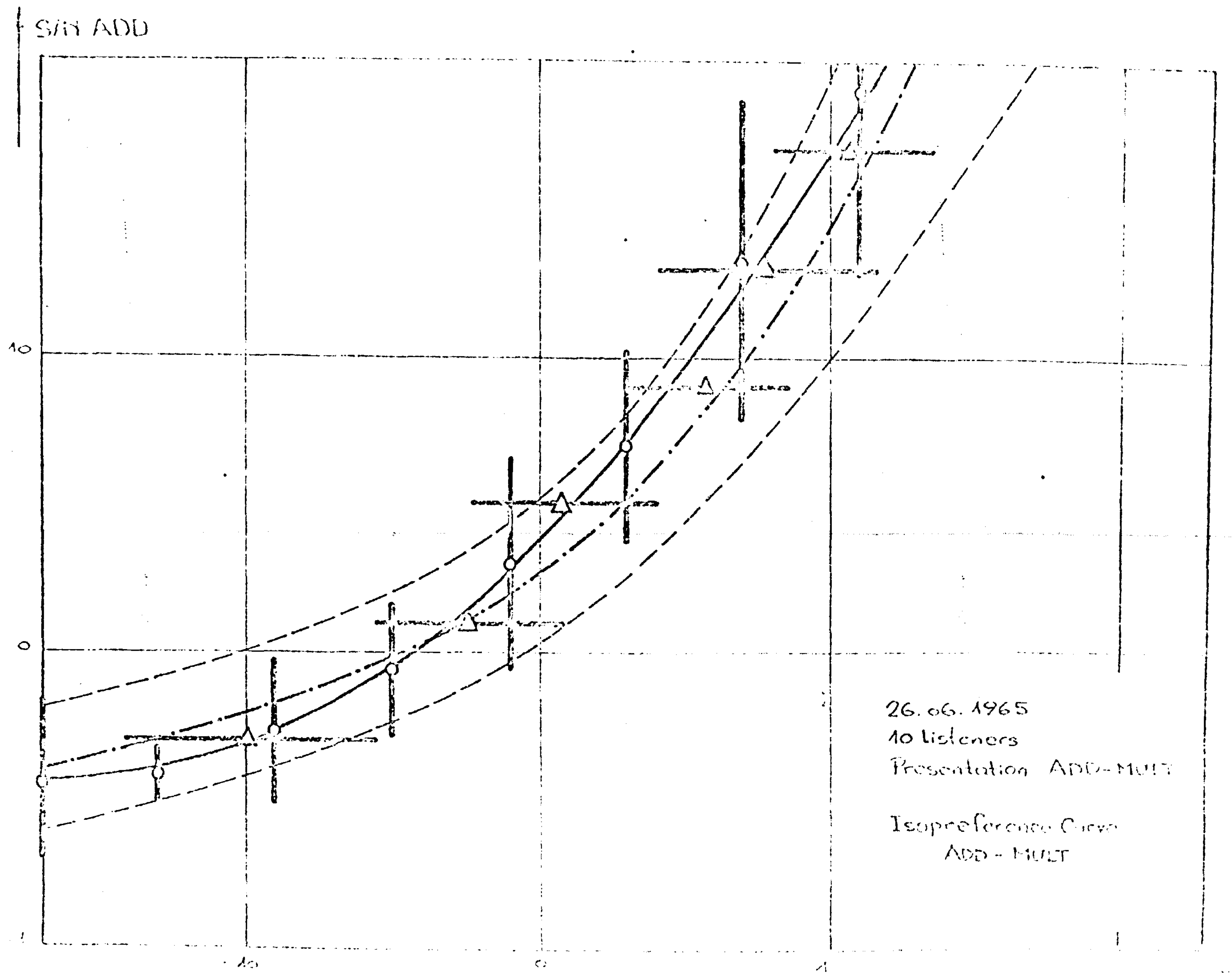


RANK-ORDERING BY PREFERENCE TESTS

ADDITIVE REFERENCE

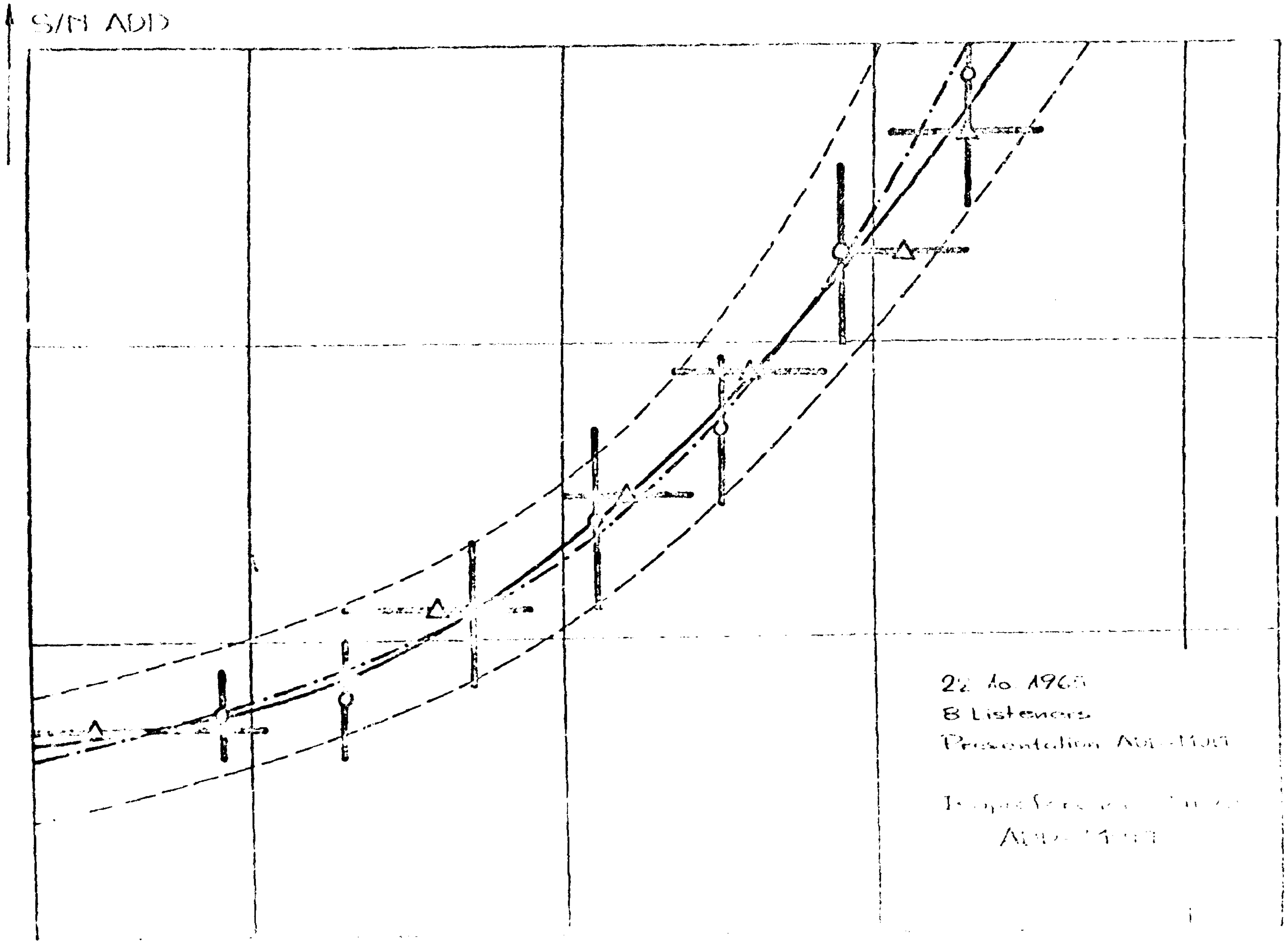


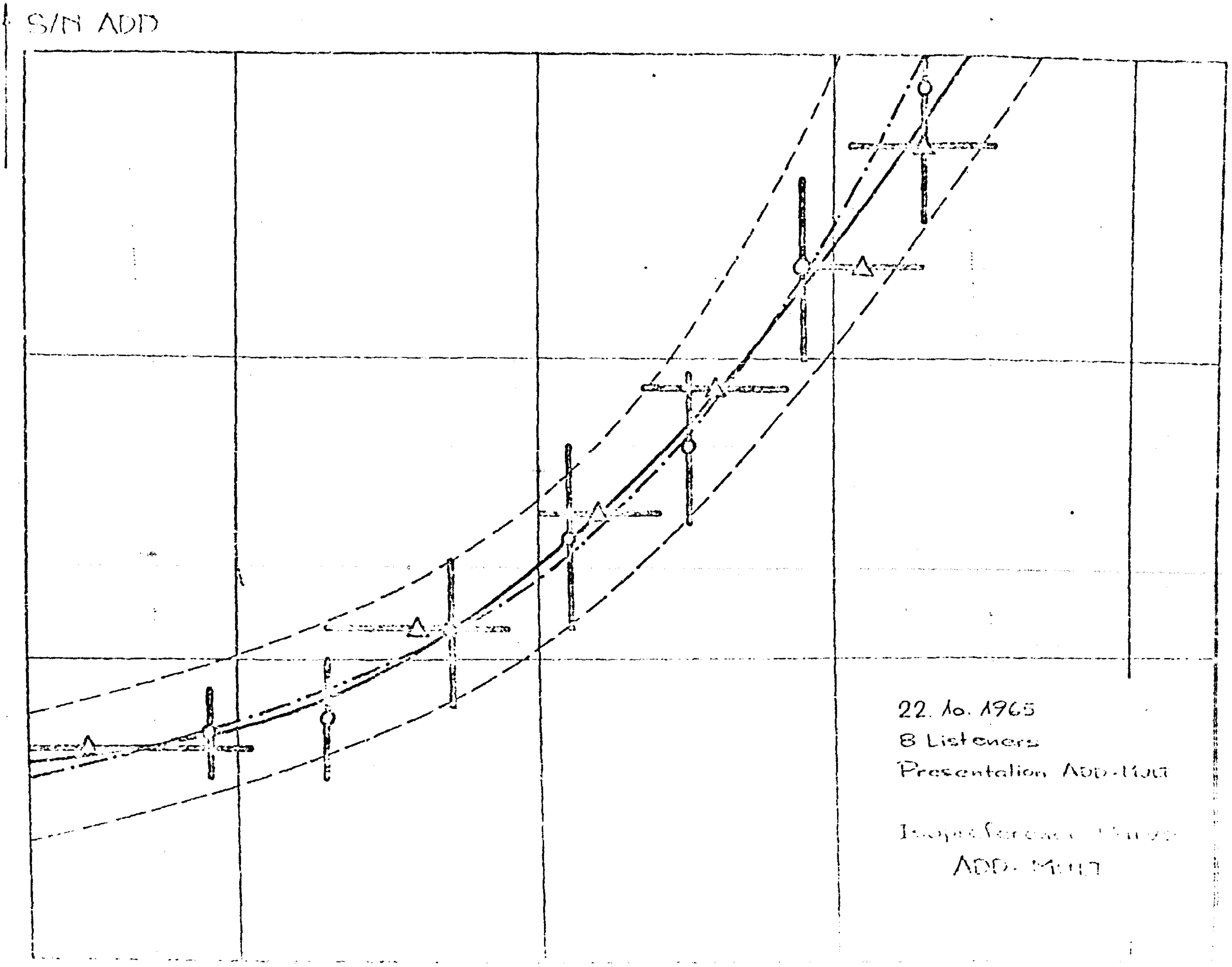
DISTRIBUTION OF ISOPREFERENCE LEVELS FOR LARGE GROUP



26.06.1965
10 listeners
Presentation ADD-MULT

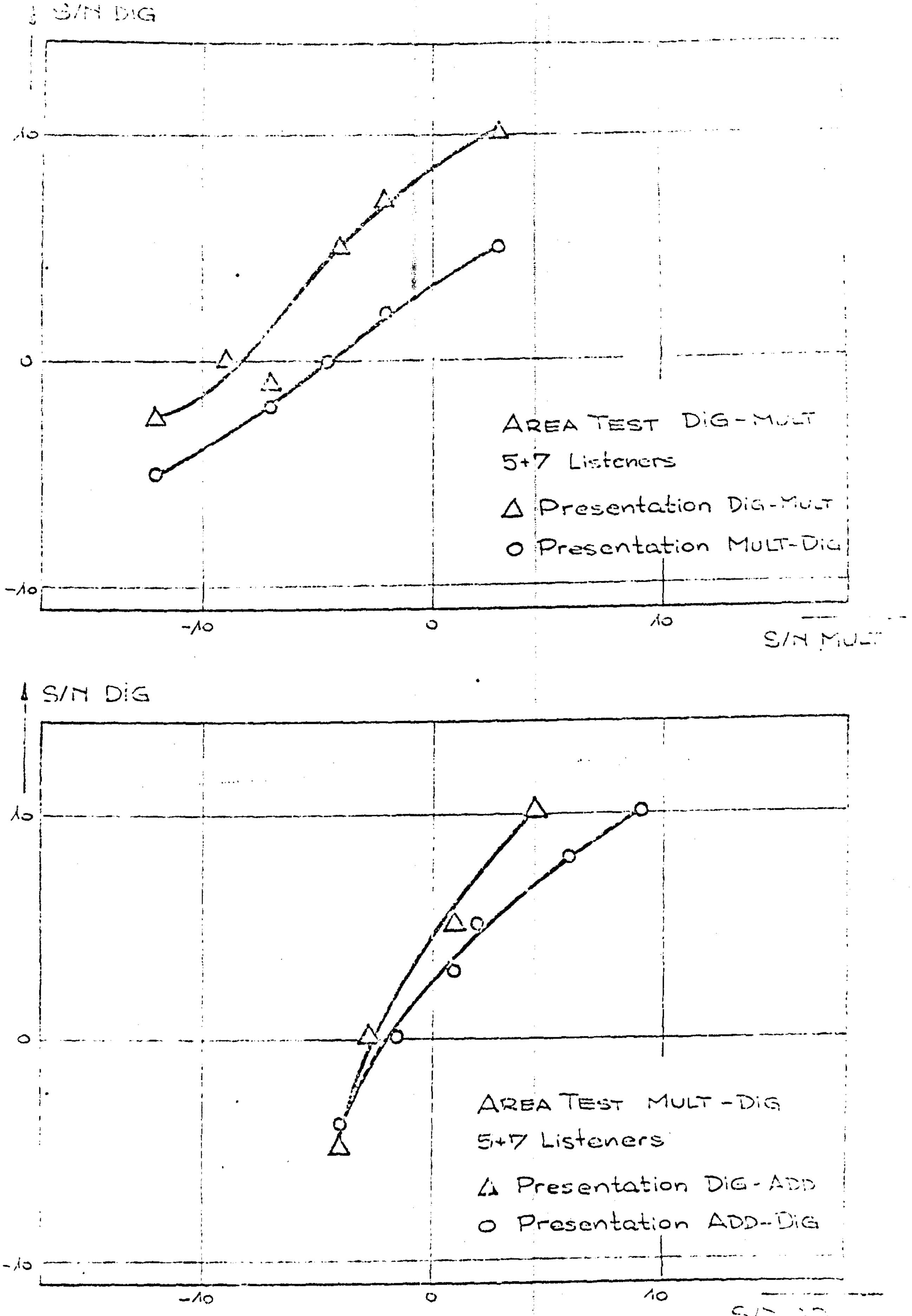
Isopreference Curve
ADD - MULT


S/M ADD
22 to 1965
8 Listeners
Presentation ADD-MOD
Interference [illegible]
ADD-MOD



22. 10. 1965
8 Listeners
Presentation ADD-MUL

Isopreference Curve
ADD- MUL



CHECK ON ORDER EFFECT

List of test performances

This page gives a survey about all tests performed over the past report period (1 January 1965 - 31 December 1965).

[illegible] Normal Preference Tests :

	Reference	MULT	ADD	DIG
Test Signals	HP	40	30	2
	LP	30	20	2
	VON	60	50	4
	VO2	20	30	2
	others	10	10	-

40 Area Tests

	MULT	ADD	DIG
MULT	-	10	2
ADD	10	-	2
DIG	2	2	-
others	6	4	2

5 Test sessions on Intelligibility

5 Test sessions on Loudness

Total 80 Test sessions 2 hours each 8 listeners in average
1300 listener hours
Approximately 64 000 decisions.

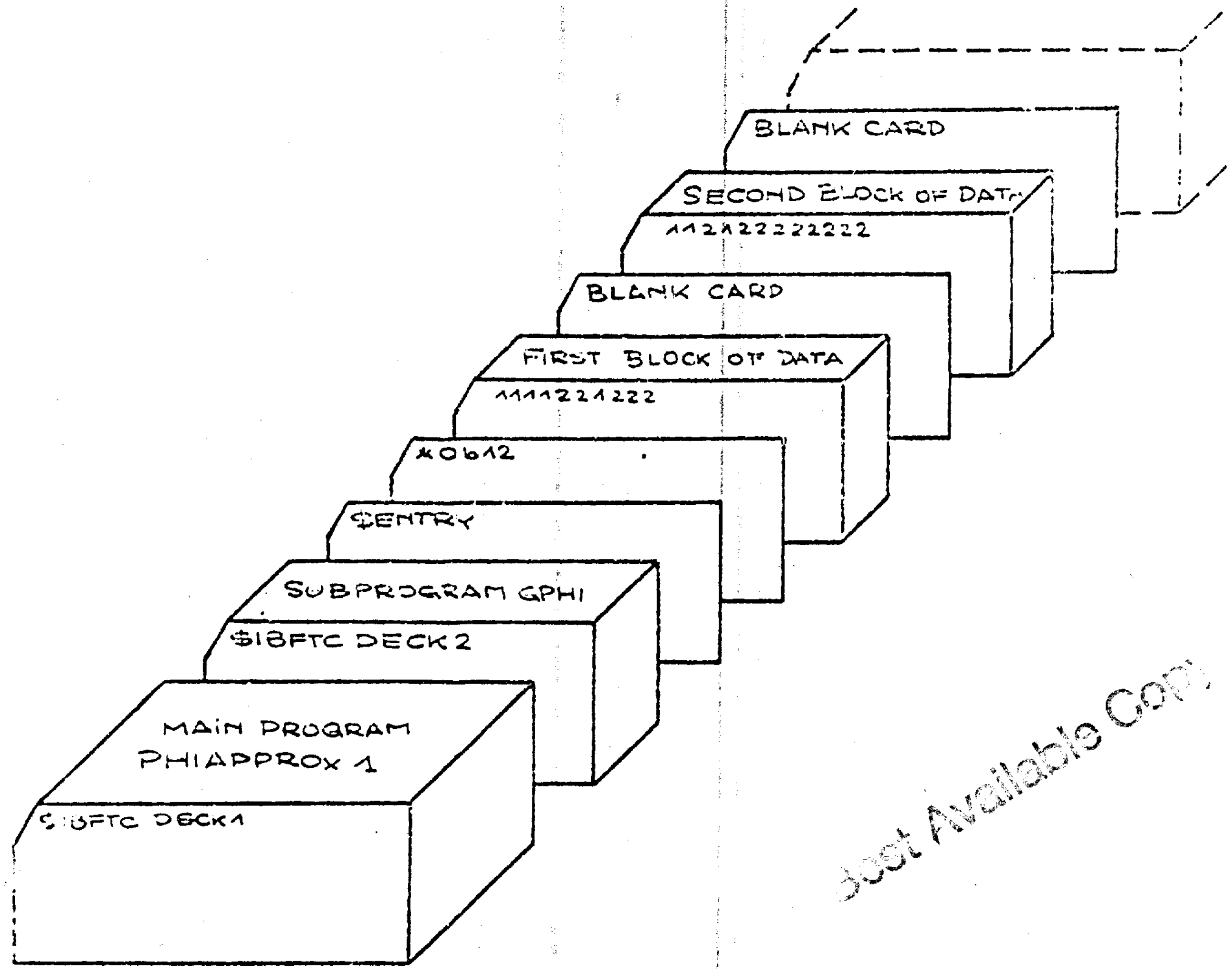
Use of the Computer Program PHIAPPROX 1

Premises :
Integer valued levels and attenuator settings have to be used. The attenuator setting must not exceed 40 dB.

The speech level has to be fixed throughout the test run.

Reference signals have to be taken from a close series of references having incremental steps of 1 dB in their S/N ratios.

Data cards have to be punched in the format given beneath thus including all information for one listener during a single test run. Cards to be evaluated together have to succeed one another thus forming a "block of data cards". End of a block of data cards is given by a card which is blank at column 80. The first block of data cards has to be preceded by a card having punched the printout symbols * O, blank, 1, 2 at the columns 1, 2, 3, 4, 5 respectively. The arrangement of cards can be seen from the following figure.

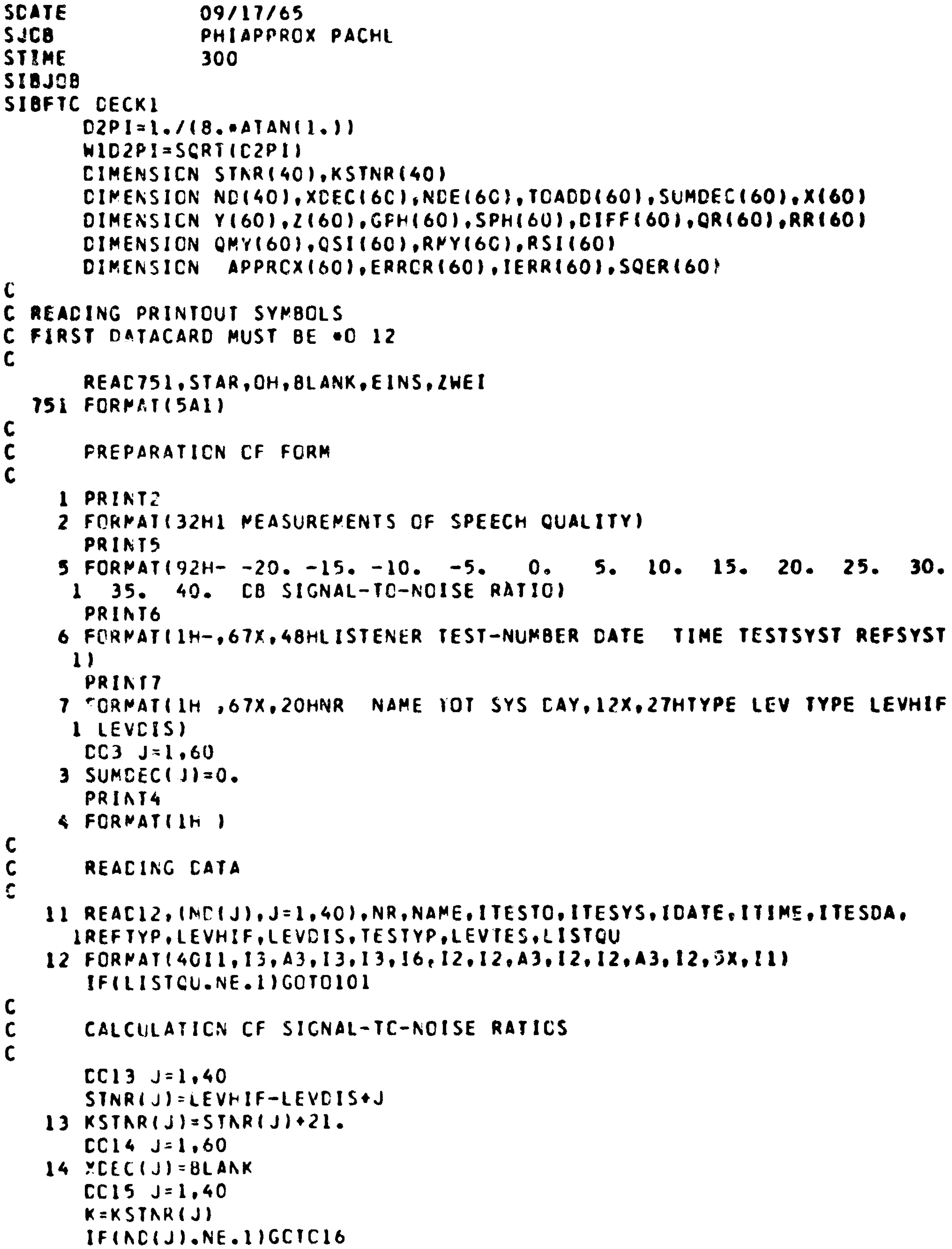


Format of Data Cards

Description	Name	Format	

Decisions of listener "1" for A and "2" for B Column number corresponds to the respective attenuator setting of the distortion signal	ND(J)
	J = 1,40
		40I1	1 through 40

Number of listener	NR	I3	41,42,43
Name of listener	NAME	A3	44,45,46
Number of test run attended totally by the listener	ITESTO	I3	47,48,49
Number of test run with this test signal attended by the listener	ITESTYS	I3	50,51,52
Date of test session DAY, MONTH, (YEAR - 1900)	IDATE	I6	53 through 58
Time of test session (HOUR)	ITIME	I2	59,60
No. of test run in this session	ITESDA	I2	61,62
Type of reference signal	REFTYP	A3	63,64,65
Level of hifi speech signal	LEVHIF	I2	66,67
Level of distortion signal	LEVDIS	I2	68,69
Type of test signal	TESTYP	A3	70,71,72
Level of test signal	LEVTES	I2	73,74
Declaration of listener card	LISTQU	I1	80



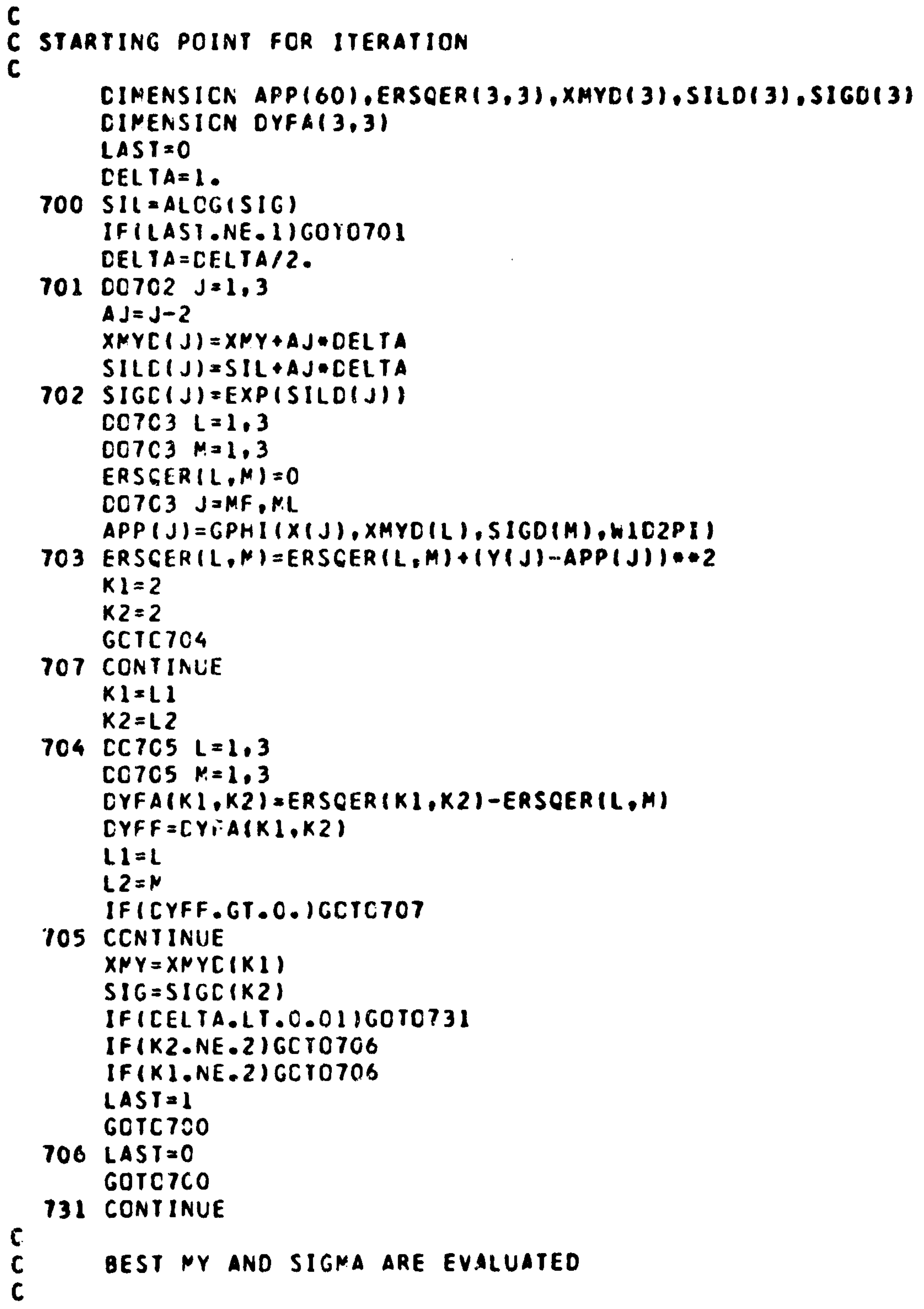
```
SCATE           09/17/65
SJCB            PHIAPPROX PACHL
STIME           300
SIBJCB
SIBFTC CECK1
      D2PI=1./(8.*ATAN(1.))
      W1D2PI=SCRT(C2PI)
      CIMENSICN STNR(40),KSTNR(40)
      CIMENSION ND(40),XCEC(6C),NDE(6C),TOADD(60),SUMDEC(60),X(60)
      DIMENSICN Y(60),Z(60),GPH(60),SPH(60),CIFF(60),QR(60),RR(60)
      CIMENSION QMY(60),QSI(60),RMY(6C),RSI(60)
      DIMENSICN  APPRCX(60),ERRCR(60),IERR(60),SQER(60)
C
C READING PRINTOUT SYMBOLS
C FIRST DATACARD MUST BE *0 12
C
      READ751,STAR,OH,BLANK,EINS,ZWEI
  751 FORMAT(5A1)
C
C     PREPARATICN CF FORM
C
    1 PRINT2
    2 FORMAT(32H1 MEASUREMENTS OF SPEECH QUALITY)
      PRINT5
    5 FORMAT(92H- -20. -15. -10.  -5.   0.   5.  10.  15.  20.  25.  30.
     1  35.  40.  CB SIGNAL-TC-NOISE RATIO)
      PRINT6
    6 FORMAT(1H-,67X,48HLISTENER TEST-NUMBER DATE  TIME TESTSYST REFSYST
     1)
      PRINT7
    7 FORMAT(1H ,67X,20HNR  NAME YOT SYS CAY,12X,27HTYPE LEV TYPE LEVHIF
     1 LEVCIS)
      CC3 J=1,60
    3 SUMCEC(J)=0.
      PRINT4
    4 FORMAT(1H )
C
C     READING CATA
C
   11 READ12,(NC(J),J=1,40),NR,NAME,ITESTO,ITESYS,IDATE,ITIME,ITESDA,
     1REFTYP,LEVHIF,LEVCIS,TESTYP,LEVTES,LISTQU
   12 FORMAT(40I1,I3,A3,I3,I3,I6,I2,I2,A3,I2,I2,A3,I2,5X,I1)
      IF(LISTCU.NE.1)GOTO101
C
C     CALCULATICN CF SIGNAL-TC-NOISE RATICS
C
      CC13 J=1,40
      STNR(J)=LEVHIF-LEVCIS+J
   13 KSTNR(J)=STNR(J)+21.
      CC14 J=1,60
   14 XCEC(J)=BLANK
      CC15 J=1,40
      K=KSTNR(J)
      IF(NC(J).NE.1)GCTC16
```

Computer Program for Data Evaluation

```
      XDEC(K)=EINS
      GOTO15
   16 IF(ND(J).NE.2)GOTO15
      XDEC(K)=ZWEI
   15 CONTINUE
      PRINT17,(XDEC(K),K=1,60),NR,NAME,ITESTO,ITESYS,ITESDA,IDATE,ITIME,
     1TESTYP,LEVTES,REFTYP,LEVHIF,LEVCIS
   17 FORMAT(1H ,4X,60A1,I6,2X,A3,2I4,I3,I7,I3,3H00 ,A3,I4,2X,A3,2I7)
C
C     PREPARING DATA FOR COMPUTATION
C
      NP=60
      DIMENSION  NOLIST(60),XNOLIS(60)
      DO21 J=1,60
   21 TOACD(J)=0
      DO22 J=1,40
      K=KSTNR(J)
      IF(ND(J).EQ.0)GOTO22
      TOACD(K)=ND(J)+999
   22 CONTINUE
      DO23 J=1,60
   23 SUMDEC(J)=SUMDEC(J)+TOACD(J)
      GOTO11
  101 CONTINUE
      DO102 J=1,60
      NOLIST(J)=SUMDEC(J)/1000.
  102 XNOLIS(J)=NOLIST(J)
      DO103 J=1,60
      MF=J
      IF(NOLIST(J).NE.0)GOTO105
  103 CONTINUE
  105 CONTINUE
      DO106 J=1,60
      K=61-J
      ML=K
      IF(NOLIST(K).NE.0)GOTO107
  106 CONTINUE
  107 CONTINUE
      MFM=MF-1
      IF(MFM.LT.1)GOTO109
      DO108 J=1,MFM
  108 Y(J)=0
  109 DO110 J=MF,ML
  110 Y(J)=(SUMDEC(J)-1000.*XNOLIS(J))/XNOLIS(J)
      MLP=ML+1
      IF(MLP.GT.60)GOTO113
      DO112 J=MLP,60
  112 Y(J)=1
  113 DO114 J=1,60
  114 X(J)=J-21
```

Computer Program for Data Evaluation

```
C
C     TEST WETHER 00001111
C
      DO42 J=1,60
      JA=J
      IF(Y(J).NE.0.)GOTO43
   42 CONTINUE
   43 DO44 J=JA,60
      IF(Y(J).NE.1.)GOTO45
   44 CONTINUE
      GOTO46
   45 GOTO48
   46 XMY=FLOAT(JA)-21.5
      SIG=0.
      PRINT4
      PRINT 302,XMY
      PRINT4
      PRINT303,SIG
      PRINT47
   47 FORMAT(10H-NO ERROR )
      GOTO1
   48 CONTINUE
C
C     ALREADY COMPUTED  X(J), Y(J), J=1,NP
C
      JA=1
  401 CONTINUE
      SUPY=0
      SUNY=0
      DO402 J=1,JA
      SUPY=SUPY+Y(J)**2
  402 JB=JA+1
      DO403 J=JB,60
  403 SUNY=SUNY+(1.-Y(J))**2
      DISCRI=SUPY-SUNY
      IF(DISCRI.GE.0.)GOTO404
      JA=JA+1
      GOTO401
  404 XMY=X(JA)
      DO406 J=1,60
      IF(Y(J).NE.0.)GOTO407
  406 X1=X(J)
  407 DO408 J=1,60
      K=61-J
      IF(Y(K).NE.1.)GOTO409
  408 X2=X(K)
  409 SIG=(X2-X1)/4.
  204 IF(SIG.GT.0.)GOTO 205
      SIG=1.
  205 CONTINUE
C
C     XMY,SIG ALREADY COMPUTED
C
```



```
C
C STARTING POINT FOR ITERATION
C
      DIMENSION APP(60),ERSQER(3,3),XMYD(3),SILD(3),SIGD(3)
      DIMENSION DYFA(3,3)
      LAST=0
      DELTA=1.
  700 SIL=ALOG(SIG)
      IF(LAST.NE.1)GOTO701
      DELTA=DELTA/2.
  701 DO702 J=1,3
      AJ=J-2
      XMYD(J)=XMY+AJ*DELTA
      SILD(J)=SIL+AJ*DELTA
  702 SIGD(J)=EXP(SILD(J))
      DO703 L=1,3
      DO703 M=1,3
      ERSQER(L,M)=0
      DO703 J=MF,ML
      APP(J)=GPHI(X(J),XMYD(L),SIGD(M),W1D2PI)
  703 ERSQER(L,M)=ERSQER(L,M)+(Y(J)-APP(J))**2
      K1=2
      K2=2
      GOTO704
  707 CONTINUE
      K1=L1
      K2=L2
  704 DO705 L=1,3
      DO705 M=1,3
      DYFA(K1,K2)=ERSQER(K1,K2)-ERSQER(L,M)
      DYFF=DYFA(K1,K2)
      L1=L
      L2=M
      IF(DYFF.GT.0.)GOTO707
  705 CONTINUE
      XMY=XMYD(K1)
      SIG=SIGD(K2)
      IF(DELTA.LT.0.01)GOTO731
      IF(K2.NE.2)GOTO706
      IF(K1.NE.2)GOTO706
      LAST=1
      GOTO700
  706 LAST=0
      GOTO700
  731 CONTINUE
C
C     BEST MY AND SIGMA ARE EVALUATED
C
```

Computer Program for Data Evaluation

```
C
C     PRINTOUT OF RESULTS
C
      PRINT 4
  301 PRINT302,XMY
  302 FORMAT(21H MEAN VALUE          =,F6.2,3H DB)
      PRINT4
      PRINT303,SIG
  303 FORMAT(21H STANDARD DEVIATION =,F6.2,3H DB)
  304 DO305 J=1,60
      APPROX(J)=GPHI(X(J),XMY,SIG,W1D2PI)
      ERROR(J)=Y(J)-APPROX(J)
  305 IERR(J)=100.*ERROR(J)+0.5
      PRINT4
      PRINT306,(IERR(J),J= 1,30)
      PRINT306,(IERR(J),J=31,60)
  306 FORMAT(7H/ERROR=,30I4)
      PRINT4
      SME=0.
      DO307 J=1,60
      SQER(J)=ERROR(J)**2
  307 SME=SME+SQER(J)/20.
      PRINT308,SME
  308 FORMAT(21H/MEAN SQUARE ERROR  =,F12.10)
      GOTO1
      END
```

Computer Program for Data Evaluation

```
SIBFTC DECK2
C
C      CALCULATION CF CUMULATIVE NCRMAL DISTRIBUTION
C
       FUNCTICN GPHI(A,B,C,D)
       Z=(A-B)/C
       IF(Z.GT.5.)GCTO3
       IF(Z.LT.(-5.))GCTO4
       Z=Z/SCRT(2.)
       ZET=Z
       Z=ABS(Z)
       A1=0.278393
       A2=C.230389
       A3=0.C00972
       A4=0.078108
       XNEN=(((A4*Z+A3)*Z+A2)*Z+A1)*Z+1.
       XNEN=XNEN**4
       ERFI=1.-1./XNEN
       IF(ZET.GE.0.)GOTO1
       GPHI=0.5-ERFI/2.
       GCTC2
     1 GPHI=C.5+ERFI/2.
     2 CONTINUE
       GCTC5
     3 GPHI=1
       GOTC5
     4 GPHI=0
     5 CONTINUE
       RETURN
       ENC
SENTRY
*0 12
      11111211111212222222                  022DRK025004210565120lHAM7188VON84      1
      1,111111111222222222                  010DRW043008210565120lHAM7188VON84      1
      11111111212222222222                  CO2MUE049013210565120lHAM7188VON84      1
      11111111112222222222                  005SVE049013210565120lHAM7188VON84      1
      1111111111'  :2222222                 C04SPA04300821056512 1HAM7188VON84      1
      11111111111111222222                  011KLE013C0321056512 1HAM7188VON84      1
      11111111122222222222                  0098ER04300821056512 1HAM7188VON84      1
      11111221122222222222                  C07PRA04901321056512 1HAM7188VON84      1
      11111222222222222222                  001LER049C1321056512 1HAM7188VON84      1
      11111111212222222222                  008ZLA043C0821056512 1HAM7188VON84      1
```

Computer Program for Data Evaluation

```
MEASUREMENTS OF SPEECH CUALITY

-20. -15. -10.  -5.   0.   5.  10.  15.  20.  25.  30.  35.  40.  DB SIGNAL-TO-NOISE RATIO

                                                                   LISTENER TEST-NUMBER CATE   TIME TESTSYST REFSYST
                                                                   NR  NAME TOT SYS CAY             TYPE LEV TYPE LEVHIF LEVDIS

                 11111211111212222222                               22  CRK  25   4   1 210565 1200 VON  84  HAM     71     88
                 11111111112222222222                               10  CRW  43   8   1 210565 1200 VON  84  HAM     71     88
                 11111111212222222222                                2  MUE  49  13   1 210565 1200 VON  84  HAM     71     88
                 11111111112222222222                                5  SVE  49  13   1 210565 1200 VON  84  HAM     71     88
                 11 11111111122222222                                4  SPA  43   8   1 210565 1200 VON  84  HAM     71     88
                 11.11111111111222222                               11  KLE  13   3   1 210565 1200 VON  84  HAM     71     88
                 11111111122222222222                                9  BER  43   8   1 210565 1200 VON  84  HAM     71     88
                 11111221122222222222                                7  PRA  49  13   1 210565 1200 VON  84  HAM     71     88
                 11111222222222222222                                1  LER  49  13   1 210565 1200 VON  84  HAM     71     88
                 11111111212222222222                                8  ZLA  43   8   1 210565 1200 VON  84  HAM     71     88

MEAN VALUE          = -1.71 CB

STANCARD CEVIATICN =  2.77 CB

ERRCR.  0  0  C  0  0  C  0  C  0  0  C  0 -0 -2 24  8 -10 -1 -15 10  7 -2 -0  5  2  1  0  0  0  0
ERRCM.  0  0  C  0  0  C  0  C  0  0  0  0  0  C  0  0  0  0   0  0  0  0  0  0  0  0  0  0  0  0

MEAN SCUARE ERROR  =  0.C0598327
```

Printout Sample